Combinatoire des mots

Julien Cassaigne
Institut de mathématiques de Luminy — CNRS
case 907, 163 avenue de Luminy, F-13288 Marseille Cedex 9, France
cassaigne@iml.univ-mrs.fr

École CIMPA, Bobo-Dioulasso, 29 octobre - 9 novembre 2012

Table des matières

1 Introduction

La combinatoire des mots, c'est l'études des suites (finies ou infinies) à valeurs dans un ensemble fini.

1.1 Historique

- Avant 1900 : quelques apparitions de la notion de mot, par exemple chez Gauss ou Prouhet.
- 1906 : Axel Thue est le premier à étudier les mots pour eux-mêmes (mots sans carrés, sans cubes).
- Première moitié du 20e siècle : quelques travaux isolés, notamment ceux de Morse (et Hedlund) et d'Arshon, avec en vue des applications à d'autres domaines des mathématiques (dynamique, etc.).
- À partir de 1960 : la théorie démarre vraiment avec l'école de Schützenberger en France et celle de Novikov et Adian en Russie.
- 1983 : parution du premier livre de M. Lothaire, *Combinatorics on words* [6].
- Depuis, la théorie a connu de nombreux développements, et de nouveaux ouvrages sont parus : les autres tomes de M. Lothaire [7, 8], les livres de N. Pytheas Fogg [11], d'Allouche et Shallit [2], le livre collectif CANT [4], etc. Voir aussi le tutoriel de Berstel et Karhumäki [3].

1.2 Définitions, notations, vocabulaire

Un *alphabet* est un ensemble fini A (selon les situations, ses éléments seront des lettres, des chiffres, des symboles, etc.). Exemples : $\{0,1\}$, $\{a,b,c\}$, $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$.

Un *mot* (fini) sur A est une suite finie d'éléments de A : $w = (a_1, \ldots, a_n)$, avec $n \in \mathbb{N}$ et $a_i \in A$ pour chaque i. La *longueur* de w est $|w| = n$. On a $|w| = \sum_{a\in A} |w|_a$, où $|w|_a$ est le *nombre d'occurrences* de a dans w.

L'ensemble des mots de longueur n sur A est noté A^n. En particulier, $A^0 = \{\varepsilon\}$ (le *mot vide*) et on identifie A^1 et A. On note $A^* = \bigcup_{n\in\mathbb{N}} A^n$ et $A^+ = \bigcup_{n\geqslant 1} A^n$.

La *concaténation* est l'opération qui à deux mots $u = (a_1, \ldots, a_m)$ et $v = (b_1, \ldots, b_n)$ associe le mot $u.v = (a_1, \ldots, a_m, b_1, \ldots, b_n)$. C'est une loi associative sur A^*, d'élément neutre ε, ce qui munit donc A^* d'une structure de monoïde, le *monoïde libre* sur A. Elle permet aussi de noter les mots de manière plus simple : $w = a_1 a_2 \ldots a_n$.

Si un mot admet une *factorisation* $w = xyz$, alors les mots x, y et z sont des *facteurs* de w. De plus, x est un *préfixe* et z un *suffixe* de w.

Un *langage* est une partie de A^*.

2 Codes et morphismes

2.1 Codes

Un langage $X \subset A^*$ est appelé *code* si tout mot de A^* admet au plus une factorisation sur X : c'est à dire si l'égalité $x_1 x_2 \ldots x_m = y_1 y_2 \ldots y_n$, avec $x_1, \ldots, x_m, y_1, \ldots, y_n \in X$, implique $m = n$ et $x_i = y_i$ pour tout i.

Codes particuliers :

- si X est un langage tel qu'aucun mot de X n'est préfixe d'un autre, alors X est un code dit *code préfixe* ;
- *code suffixe* : similaire ;
- un code à la fois préfixe et suffixe est dit *bifixe*.

Exemples : $\{a, ba\}$ est un code préfixe ; $\{abb, aab\}$ est un code bifixe ; $\{a, aba\}$ est un code qui n'est ni préfixe, ni suffixe (mais à délai de déchiffrage borné) ; $\{aa, ba, baa, bb, bba\}$ est aussi un code qui n'est ni préfixe, ni suffixe (à délai de déchiffrage non borné).

2.2 Morphismes

Une application $f\colon A^* \to B^*$ est un *morphisme de monoïdes libres* si, pour tous u et v dans A^*, on a : $f(uv) = f(u)f(v)$.

Proposition 2.1. *Un morphisme $f\colon A^* \to B^*$ est entièrement défini par sa restriction à A. Réciproquement, étant donné $g\colon A \to B^*$ une application quelconque, il existe un unique morphisme $f\colon A^* \to B^*$ qui étend g.*

Exemples classiques, pour $A = B = \{a, b\}$: le morphisme de Thue-Morse θ est défini par $\theta(a) = ab$ et $\theta(b) = ba$; le morphisme de Fibonacci φ est défini par $\varphi(a) = ab$ et $\varphi(b) = a$. Ainsi $\theta(aaba) = ababbaab$ et $\varphi(aabab) = ababaaba$.

Un morphisme f est dit *non effaçant* si $f(x) \neq \varepsilon$ pour tout $x \in A$, et *effaçant* dans le cas contraire ; *lettre-à-lettre* si $f(x) \in B$ pour tout $x \in A$; *uniforme* si tous les mots $f(x)$ pour $x \in A$ ont la même longueur.

Proposition 2.2. *Un morphisme $f\colon A^* \to B^*$ est injectif si et seulement si la restriction de f à A est injective et $f(A)$ est un code.*

Preuve. Si $f(A)$ n'est pas un code, on peut trouver un mot qui a deux factorisations distinctes $w = f(a_1)\dots f(a_m) = f(b_1)\dots f(b_n)$. Soient $x = a_1 \dots a_m$ et $y = b_1 \dots b_n$: on a $x \neq y$ et $f(x) = f(y)$, donc f n'est pas injectif.

Réciproquement, si f n'est pas injectif, soient $x = a_1 \dots a_m$ et $y = b_1 \dots b_n$ tels que $x \neq y$ et $f(x) = f(y)$. Si $m = n$ et $f(a_i) = f(b_i)$ pour tout i, il existe i tel que $a_i \neq b_i$ donc la restriction de f à A n'est pas injective. Sinon $f(A)$ n'est pas un code. □

2.3 Le théorème du défaut

Théorème 2.1 (Théorème du défaut)**.**

1. Soit X un langage fini de A^*. Si X n'est pas un code, alors il existe un langage Y de cardinal strictement inférieur à celui de X tel que tout mot de X se factorise sur Y.

2. Soit $f\colon A^* \to B^*$ un morphisme. Si f n'est pas injectif, alors il existe A' de cardinal strictement inférieur à celui de A, et des morphismes $g\colon A'^* \to B^*$ et $h\colon A^* \to A'^*$ tels que $f = g \circ h$.

Corollaire 2.1. *Soient $x, y \in A^+$. Les propriétés suivantes sont équivalentes :*

(i) $xy = yx$;

(ii) $\exists m \geqslant 1, \exists n \geqslant 1, x^n = y^m$;

(iii) $\{x, y\}$ *n'est pas un code ;*

(iv) $\exists z \in A^+, \{x, y\} \subset z^+ = \{z^n : n \geqslant 1\}$.

Preuve du corollaire. (i) implique (iii) par définition ; de même (ii) implique (iii) ; (iii) implique (iv) par le théorème du défaut ; (iv) implique (i) car $z^m z^n = z^n z^m$; (iv) implique (ii) car $(z^m)^n = (z^n)^m$. □

Preuve du théorème. On prouve la seconde formulation.

Observons d'abord que si f est effaçant, c'est à dire s'il existe $a \in A$ tel que $f(a) = \varepsilon$, le résultat est clair : il suffit de prendre $A' = A \setminus \{a\}$, g restriction de f à A', et h qui efface les a en laissant les autres lettres inchangées.

La preuve est par récurrence sur $|f| = \sum_{a \in A} |f(a)|$. Si f est non injectif et non effaçant, alors il existe deux mots $u \neq v$ tels que $f(u) = f(v)$. Quitte à raccourcir et échanger u et v, on peut supposer que leurs premières lettres respectives a et b sont distinctes et que $f(a)$ est préfixe de $f(b)$: $f(b) = f(a)z$. Soit alors h_0 et f' les morphismes définis par $h_0(b) = ab$, $f'(b) = z$, $h_0(x) = x$ et $f'(x) = f(x)$ si $x \neq b$. Alors $f = f' \circ h_0$ et $|f'| < |f|$. Comme h_0 est injectif (en effet $h_0(A)$ est un code suffixe), f' n'est pas injectif, et par hypothèse de récurrence f' s'écrit $f' = g \circ h'$. On pose alors $h = h' \circ h_0$ pour obtenir $f = g \circ h$. □

3 Conjugaison, périodicité, répétitions

3.1 Racine primitive

Soit $w \in A^*$. On dit que w est *primitif* si $w \neq \varepsilon$ et w n'est pas une puissance d'un autre mot. Ainsi, w est primitif si et seulement si

$$\forall x \in A^*, \forall n \in \mathbb{N}, (w = x^n \Longrightarrow n = 1).$$

Proposition 3.1. *Tout mot w non vide s'écrit de manière unique $w = x^n$, avec x primitif et $n \in \mathbb{N}$.*

Le mot x est appelé *racine primitive* de w.

Preuve. Existence : si w est primitif, $w = w^1$; sinon, $w = w'^k$ avec $k \geqslant 2$ et on conclut par récurrence sur $|w|$.

Unicité : si $w = x^m = y^n$, on applique le corollaire 2.1 ((ii) implique (iv)) pour conclure que x et y ne peuvent être distincts et primitifs. □

Application : pour $x \in A^+$, cherchons à déterminer l'ensemble $C(x)$ des mots de A^* qui commutent avec x.

Proposition 3.2. *$C(x) = z^* = \{z^n : n \in \mathbb{N}\}$ où z est la racine primitive de x.*

Preuve. Clairement $z^* \subset C(x)$; et $C(x) \subset z^*$ par le corollaire 2.1 ((i) implique (iv)). □

3.2 Période, répétitions

On dit que x est une *période* d'un mot w s'il existe $n \in \mathbb{N}$ tel que w est préfixe de x^n.

Par exemple, *abacaba* a pour périodes *abac*, *abacab*, mais aussi *abacaba*, *abacababcd*, etc. : la définition n'interdit pas $|x| > |w|$.

Soit $p \in \mathbb{N}$: on dit que w est *p-périodique* s'il a une période x telle que $|x| = p$.

Proposition 3.3. *Soit $x \in A^+$. Le mot x est une période de w si et seulement si w est préfixe de xw.*

Preuve. Si w est préfixe de x^n, alors xw est préfixe de x^{n+1} donc w et xw sont préfixes du même mot. Comme $|w| < |xw|$, w est préfixe de xw.

Réciproquement, si w est préfixe de xw, on en déduit (par récurrence sur n) que w est préfixe de $x^n w$ pour tout $n \in \mathbb{N}$. Pour n assez grand w est donc préfixe de x^n (ici on utilise l'hypothèse $x \neq \varepsilon$). □

Soit α un rationnel positif et x un mot tel que $\alpha|x|$ est un entier. Il existe un unique mot w de longueur $\alpha|x|$ tel que x est période de w. On le note $w = x^\alpha$, et on l'appelle *répétition* (ou *puissance fractionnaire*) de x d'exposant α.

Par exemple, $(abb)^{8/3} = abbabbab$. Attention, $x^\alpha x^\beta$ est en général différent de $x^{\alpha+\beta}$.

3.3 Conjugaison

Deux mots x et y de A^* sont dits *conjugués* s'il existe $z \in A^*$ tel que $xz = zy$. Ils sont dits *transposés* s'il existe u et v tels que $x = uv$ et $y = vu$.

Proposition 3.4. *Deux mots sont conjugués si et seulement s'ils sont transposés. On définit ainsi une relation d'équivalence $\sim$ sur A^*.*

Preuve. Si $x = uv$ et $y = vu$, alors $xu = uy$. Réciproquement, supposons que $xz = zy$. Si x est vide, y aussi donc x et y sont transposés. Sinon, x est une période de z par la proposition 3.3. Soit $n = \lfloor |z|/|x| \rfloor$: z peut s'écrire $z = x^n u$ avec u préfixe de x. Soit v tel que $x = uv$. Alors $zy = xz = x^{n+1}u = x^n uvu = zvu$, donc $y = vu$.

Relation d'équivalence : exercice facile. Utiliser les deux définitions (transposition pour la symétrie, conjuguaison pour la transitivité). □

Remarque. Pourquoi deux noms pour la même notion ? Ces deux définitions sont en fait valables dans n'importe quel monoïde, mais elles ne sont en général pas équivalentes : deux éléments peuvent être conjugués sans être transposés. Dans le monoïde libre, elles sont équivalentes, et on emploie plutôt le terme "conjugués".

Exercice 3.1. Montrer qu'un mot w est primitif si et seulement s'il a exactement $|w|$ conjugués distincts (y compris lui-même).

3.4 Le théorème de Fine et Wilf

Théorème 3.1 (Théorème de Fine et Wilf)**.**

1. Soient x et y deux périodes distinctes de $w \in A^+$. Si $|w| \geqslant |x| + |y| - \mathrm{pgcd}(|x|, |y|)$, alors x et y ont la même racine primitive, qui est alors une période de x dont la longueur divise $\mathrm{pgcd}(|x|, |y|)$.

2. Soient $w \in A^+$ et p, q des entiers strictement positifs. Si w est à la fois p-périodique et q-périodique, et si $|w| \geqslant p + q - \mathrm{pgcd}(p, q)$, alors w est $\mathrm{pgcd}(p, q)$*-périodique.*

Preuve. Prouvons la première formulation (la seconde s'en déduit). Soient $p = |x|$, $q = |y|$, $d = \mathrm{pgcd}(p, q)$. On suppose, quitte à remplacer w par un préfixe, que $|w| = p + q - d$.

Supposons d'abord que $d = 1$. On sait que w est préfixe de xw et yw, donc aussi de xyw et de yxw. Comme $|w| = p+q-1$, les mots xy et yx ne peuvent différer que par leur dernière lettre : posons $xy = wa$ et $yx = wb$ avec $a, b \in A$. Comme $|xy|_a = |x|_a + |y|_a = |yx|_a$, on a $|wa|_a = |wb|_a$ et donc $b = a$. Par conséquent $xy = yx$, et par le corollaire 2.1 x et y sont puissances d'un même mot z. La longueur de z divise p et q, donc $|z| = 1$: z est une lettre et les mots x, y et w sont donc constants.

On se ramène maintenant au cas $d = 1$. Soit $B = A^d$, que l'on traite comme un alphabet. On définit un morphisme $f\colon B^* \to A^*$ en posant $f(b) = b$ si $b \in B$: ce morphisme est injectif car B est un code (bifixe), et $f(B^*)$ est le langage des mots dont la longueur est multiple de d. On pose alors $x' = f^{-1}(x)$, $y' = f^{-1}(y)$, $w' = f^{-1}(w)$: il est clair que x' et y' sont des périodes distinctes de w', de longueurs $p' = p/d$ et $q' = q/d$ premières entre elles, et on a $|w'| = p' + q' - 1$. D'après le paragraphe précédent, x' et y' sont puissances d'une même lettre $z' \in B$. Alors x et y sont puissances de $f(z')$, donc ils ont la même racine primitive qui est aussi la racine primitive de $f(z')$, et dont la longueur divise donc d. □

Exercice 3.2. Prouver l'optimalité du théorème de Fine et Wilf : pour tous p et q tels que $d = \mathrm{pgcd}(p, q) < p < q$, il existe un mot w de longueur $p + q - d - 1$ qui est p-périodique et q-périodique, mais qui n'est pas d-périodique. Commencer par le cas $d = 1$, et utiliser l'algorithme d'Euclide.

4 Construction de mots infinis

4.1 Définitions

Un *mot infini* est une suite $\mathbf{u} = (u_n)_{n \in \mathbb{N}}$ indexée par les entiers naturels et à valeurs dans un alphabet A. On note $A^{\mathbb{N}}$ (ou A^ω) l'ensemble des mots infinis sur A.

Si $w \in A^*$ et $\mathbf{u} \in A^{\mathbb{N}}$, on définit $\mathbf{v} = w\mathbf{u} \in A^{\mathbb{N}}$ en insérant w au début de $\mathbf{u}$: ainsi $v_0 v_1 \dots v_{|w|-1} = w$, et $v_{|w|+i} = u_i$ pour tout $i \in \mathbb{N}$. Cela permet de définir naturellement les notions de facteur (fini), préfixe (fini) et suffixe (infini) d'un mot infini. Si w est facteur de $\mathbf{u}$, il existe un entier $i \in \mathbb{N}$ pour lequel $w = u_i u_{i+1} \dots u_{i+|w|-1}$. On dit alors que w *apparaît à la position* i dans $\mathbf{u}$, ou encore que i est une *occurrence* de w dans $\mathbf{u}$. On note $F_n(\mathbf{u})$ le langage des facteurs de longueur n de $\mathbf{u}$, et $F(\mathbf{u})$ le langage de tous les facteurs de $\mathbf{u}$, appelé aussi *langage de* $\mathbf{u}$.

Un mot infini $\mathbf{u}$ est dit *p-périodique*, ou simplement *périodique*, s'il existe $p \geqslant 1$ tel que $u_{n+p} = u_n$ pour tout $n \in \mathbb{N}$, et *ultimement périodique* si ceci n'est vrai que pour n assez grand. Un mot infini p-périodique $\mathbf{u}$ est entièrement défini par son préfixe v de longueur p : on note alors $\mathbf{u} = v^\omega$. Tout mot infini ultimement périodique peut se mettre sous la forme tv^ω (le mot fini t est alors appelé *partie transiente* ou *pré-période* de $\mathbf{u}$).

4.2 Topologie

Soit $q \in \mathbb{R}$, $q > 1$ (par exemple, $q = 2$). Pour $\mathbf{x}, \mathbf{y} \in A^{\mathbb{N}}$, on pose $d(\mathbf{x}, \mathbf{y}) = 0$ si $\mathbf{x} = \mathbf{y}$, et $d(\mathbf{x}, \mathbf{y}) = q^{-|w|}$ sinon, où w est le plus grand préfixe commun de $\mathbf{x}$ et $\mathbf{y}$. Ceci définit une distance ultramétrique sur $A^{\mathbb{N}}$, qui est donc muni d'une structure d'espace métrique.

Cette distance s'étend naturellement à $A^\infty = A^{\mathbb{N}} \cup A^*$, avec la même définition.

Proposition 4.1 (Lemme de König)**.** *Si X est un langage infini de A^*, alors l'adhérence de X dans A^∞ contient un mot infini.*

Preuve. Pour $w \in A^*$, notons X_w l'ensemble des éléments de X dont w est préfixe. On construit par récurrence un mot infini $\mathbf{u}$ ayant la propriété : pour tout $n \in \mathbb{N}$, $X_{u_0 u_1 \dots u_{n-1}}$ est infini.

Cette propriété est triviale pour $n = 0$ puisque X est infini. Si elle est vraie pour un certain n, soit $w = u_0 u_1 \dots u_{n-1}$. L'ensemble $X_w = \bigcup_{a \in A} X_{wa}$ est infini par hypothèse de récurrence, donc au moins l'un des X_{wa} est infini. On choisit un tel a et on pose $u_n = a$.

Le mot infini $\mathbf{u}$ ainsi construit est bien dans l'adhérence de X, puisque X contient pour tout n des mots v tels que $d(v, \mathbf{u}) \leqslant q^{-n}$: il suffit de prendre v dans $X_{u_0 u_1 \dots u_{n-1}}$. □

Proposition 4.2. *Les espaces $A^{\mathbb{N}}$ et A^∞ sont compacts.*

Preuve. Soit $(\mathbf{u}_n)$ une suite à valeurs dans $A^{\mathbb{N}}$. Soit v_n le préfixe de longueur n de $\mathbf{u}_n$ (attention, ici v_n est un mot et non une lettre). On applique le lemme de König à l'ensemble $X = \{v_n : n \in \mathbb{N}\}$. Le mot infini $\mathbf{v}$ obtenu est une valeur d'adhérence de $(\mathbf{u}_n)$.

Le cas de A^∞ est similaire, une valeur d'adhérence pouvant aussi être un mot fini (si ce mot apparaît une infinité de fois dans la suite). □

Remarque. L'hypothèse qu'un alphabet est toujours fini est essentielle pour les deux propositions ci-dessus.

Exercice 4.1. Montrer que A^∞ est le complété de A^*. On peut donc construire A^∞ à partir de A^* comme on construit $\mathbb{R}$ à partir de $\mathbb{Q}$.

L'espace $A^{\mathbb{N}}$ peut être muni de l'opérateur de *décalage* S, défini par $S(\mathbf{u}) = \mathbf{v}$ où $v_n = u_{n+1}$ (on efface la première lettre). Ainsi $(A^{\mathbb{N}}, S)$ est un système dynamique topologique : ceci est à l'origine du lien entre combinatoire des mots et dynamique discrète, qui sera étudié dans d'autres cours.

4.3 Mots infinis définis par morphisme

Soit $f : A^* \to B^*$ un morphisme. Alors f s'étend en une application $\tilde{f} : A^\infty \to B^\infty$ en posant

$$\tilde{f}(\mathbf{u}) = \lim_{n \to \infty} f(u_0 u_1 \dots u_{n-1}).$$

Si f est non effaçant, $\tilde{f}$ est continue (et même 1-lipschitzienne) et l'image d'un mot infini est un mot infini. Par la suite, on note f au lieu de $\tilde{f}$. Une application f ainsi obtenue est appelée *substitution*.

Soit maintenant f un endomorphisme non effaçant de A^*. On suppose de plus qu'il existe $a \in A$ telle que $f(a) \in aA^+$: un tel f est dit *itérable à partir de a*.

Proposition 4.3. *Si f est itérable à partir de a, alors la suite $(f^k(a))_{k \in \mathbb{N}}$, où $f^k = f \circ f \circ \dots \circ f$ est la composée de k copies de f, a une limite dans $A^{\mathbb{N}}$ notée $f^\omega(a)$. C'est un point fixe de f : $f(f^\omega(a)) = f^\omega(a)$.*

Preuve. On montre par récurrence que $f^k(a)$ est un préfixe de $f^{k+1}(a)$, et que $|f^k(a)| > k$. □

Exercice 4.2. Soit f un endomorphisme non effaçant de A^* (pas nécessairement itérable). Décrire tous les points fixes de f dans A^∞. Il peut y en avoir un, plusieurs, une infinité, ou même aucun.

Exemples classiques (voir section 2.2) : $\theta^\omega(a)$, dit *mot de Thue-Morse*, et $\varphi^\omega(a)$, dit *mot de Fibonacci*, sont des mots infinis binaires, points fixes respectivement de θ et φ. Ils ne sont pas ultimement périodiques (exercice ; procéder par l'absurde).

Un mot infini $\mathbf{u}$ qui peut se mettre sous la forme $\mathbf{u} = f^\omega(a)$ est dit *purement morphique*. Un mot infini de la forme $\mathbf{v} = g(\mathbf{u})$, où g est un morphisme et $\mathbf{u}$ est purement morphique, est dit *morphique*.

Exercice 4.3. Construire un mot morphique qui n'est pas purement morphique. Chercher d'abord un exemple ultimement périodique, puis un autre qui ne soit pas ultimement périodique.

4.4 Récurrence

Un mot infini est dit *récurrent* si tous ses facteurs ont une infinité d'occurrences.

Proposition 4.4. *Un mot infini* $\mathbf{u}$ *est récurrent si et seulement si une infinité de préfixes de* $\mathbf{u}$ *ont au moins une seconde occurrence dans* $\mathbf{u}$.

Preuve. Supposons que $\mathbf{u}$ satisfait la propriété en question mais n'est pas récurrent. Il existe alors un facteur w qui n'a qu'un nombre fini d'occurrences. Soit k la dernière occurrence de w dans $\mathbf{u}$. Par hypothèse, il existe un préfixe p de $\mathbf{u}$ de longueur supérieure à $k + |w|$ qui apparaît une seconde fois, disons à une position $i \geqslant 1$. Mais alors w apparaît aussi à la position $i + k$, une contradiction. La réciproque est triviale. □

Un mot infini est dit *uniformément récurrent* s'il est récurrent et si, de plus, pour tout facteur w la différence entre deux occurrences consécutives de w est bornée (on dit alors que w apparaît dans $\mathbf{u}$ *à lacunes bornées*).

La notion de récurrence uniforme est importante, car elle correspond à la notion de minimalité pour les systèmes dynamiques discrets.

5 Complexité en facteurs

Étant donné un mot infini, on peut étudier le langage de ses facteurs finis. Intuitivement, on s'attend à ce que des mots "simples" (car ils sont engendrés par des mécanismes simples, ou ont une certaine propriété de régularité, par exemple) aient aussi un langage des facteurs "simple". Une manière de quantifier cela est simplement de compter les facteurs de chaque longueur. Ce faisant, on associe au mot infini considéré une fonction : sa *fonction de complexité en facteurs*.

Cette fonction a été introduite en 1938 par Gustav A. Hedlund et Marston Morse [9], sous le nom "block growth", comme un outil pour étudier les systèmes dynamiques symboliques. Andrzej Ehrenfeucht, K. P. Lee et Grzegorz Rozenberg [5] lui donnent en 1975 le nom "subword complexity", que nous traduisons par "complexité en facteurs" (attention, en français "sous-mot" a un sens différent de "subword").

La complexité en facteurs ne doit pas être confondue avec d'autres notions de complexité, comme la complexité algorithmique ou la complexité de Kolmogorov. Nous ne les utiliserons pas ici, et quand nous écrivons "complexité" cela signifie toujours "complexité en facteurs".

5.1 Définition, propriétés de base et premiers exemples

5.1.1 Définition

On considère un mot infini $\mathbf{u} = u_0u_1u_2\ldots \in A^{\mathbb{N}}$. (On pourrait aussi travailler avec des mots bi-infinis dans $A^{\mathbb{Z}}$; la plupart des résultats que nous mentionnerons pour les mots infinis sont aussi valables pour les mots bi-infinis.)

Chacun des ensembles $F_n(\mathbf{u}) = F(\mathbf{u}) \cap A^n$ des facteurs d'une longueur donnée n de $\mathbf{u}$ est fini. Pour tout $n \in \mathbb{N}$, soit $p_{\mathbf{u}}(n)$ le cardinal de $F_n(\mathbf{u})$. Ainsi $p_{\mathbf{u}}$ est une fonction de $\mathbb{N}$ dans $\mathbb{N}$, qui est appelée *fonction de complexité* de $\mathbf{u}$. Souvent, quand aucune confusion n'est possible, on note simplement p, la référence au mot $\mathbf{u}$ restant implicite.

5.1.2 Complexité des mots périodiques

Exemple 5.1. Soit $\mathbf{u} = (aab)^\omega = aabaabaab\ldots$ Alors $F(\mathbf{u}) = (aab)^*\{\varepsilon, a, aa\} \cup (aba)^*\{\varepsilon, a, ab\} \cup (baa)^*\{\varepsilon, b, ba\}$, et la fonction de complexité de $\mathbf{u}$ est donnée par $p(0) = 1$ ($F_0(\mathbf{u}) = \{\varepsilon\}$), $p(1) = 2$ ($F_1(\mathbf{u}) = \{a, b\}$) et $p(n) = 2$ pour $n \geqslant 2$ ($F_n(\mathbf{u})$ contient exactement un mot commençant par aa, un mot commençant par ab, et un mot commençant par ba).

Plus généralement, soit $z \in A^+$ un mot primitif. Alors $\mathbf{u} = z^\omega = zzz\ldots$ est un mot infini périodique, de période minimale z. La fonction de complexité de $\mathbf{u}$ vaut $p(n) = |z|$ pour tout $n \geqslant |z|$, puisqu'il y a exactement un facteur de longueur n commençant par chaque conjugué de z.

Exercice 5.1. Soit $\mathbf{u} = z^\omega$ avec z primitif comme ci-dessus. Montrer que $p(n) = |z|$ est vrai aussi pour $n = |z| - 1$. Sur un alphabet binaire, donner un exemple avec $|z|$ arbitrairement grand pour lequel $p(|z| - 2) \neq |z|$ (facile), et un autre pour lequel $p(n) = |z|$ est vrai pour tout $n \geqslant \log_2 |z|$ (difficile).

Cette propriété s'étend aux mots ultimement périodiques.

Proposition 5.1. *Soit $\mathbf{u} = tz^\omega$ un mot infini ultimement périodique. Alors sa complexité est bornée par $|tz|$. De plus, si z est primitif et t soit est vide, soit se termine par une lettre différente de la dernière lettre de z (tout mot ultimement périodique peut s'écrire de manière unique sous cette forme), alors $p(n) = |tz|$ pour tout $n \geqslant |tz|$.*

Preuve. Si $i \geqslant |tz|$ et $n \in \mathbb{N}$, alors le facteur de longueur n de $\mathbf{u}$ qui apparaît à la position i apparaît aussi à la position $i - |z|$. Par conséquent tous les facteurs apparaissent au moins une fois à une position inférieure à $|tz|$, et comme il y a $|tz|$ telles positions, il y a au plus $|tz|$ facteurs de chaque longueur.

Supposons maintenant que z est primitif, t minimal, et $n \geqslant |tz|$. Si $p(n) < |tz|$, alors il existe deux positions $i < j < |tz|$ telles que le même facteur $w = w_0 w_1 \ldots w_{n-1}$ de longueur n apparaît aux positions i et j. Le suffixe de longueur $|z|$ de w apparaît dans $\mathbf{u}$ aux positions $i + n - |z|$ et $j + n - |z|$. Comme ces deux positions sont au moins $|t|$ et z est primitif, ce n'est possible que si $(j - i)$ est un multiple de $|z|$. Soit $j - i = m|z|$ avec $m \geqslant 1$. Par conséquent, $i \leqslant j - |z| < |t|$ et t est donc non vide. Ainsi $w_{|t|-i-1} = u_{|t|-1}$ est la dernière lettre de t. Mais $w_{|t|-i-1}$ est aussi $u_{|t|-1+m|z|}$, la dernière lettre de z, une contradiction. □

5.1.3 Complexité maximale

Soit $k = \#A$ le cardinal de A. Comme il y a k^n mots dans A^n, clairement $p(n) \leqslant k^n$. Cette borne peut être atteinte, comme le montre la construction suivante.

Exemple 5.2. Le monoïde libre A^* est un ensemble dénombrable. Soit $A^* = \{w_0, w_1, w_2, \ldots\}$ une énumération de A^*, par exemple dans l'ordre généalogique : $A^* = \{\varepsilon, a, b, aa, ab, ba, bb, aaa, aab, \ldots\}$ (pour un alphabet binaire). Soit $\mathbf{u} = w_0 w_1 w_2 \ldots$ le mot infini obtenu en concaténant les w_i :

$$\mathbf{u} = abaaabbabbaaaaababaabbbaababbbabbbaaaaaaaabaaba\ldots$$

Ce mot est appelé *mot de Champernowne* et sa complexité est $p(n) = k^n$, puisque par construction tout mot de A^* apparaît dans $\mathbf{u}$.

Exercice 5.2. Soit L le langage rationnel $L = \{a, ba\}^*$ (en d'autres termes, les mots qui ne contiennent pas bb et ne se terminent pas par b). Définir un mot infini en concaténant les éléments de L dans n'importe quel ordre. Montrer que la complexité du mot obtenu est indépendante de l'ordre d'énumération de L, et la calculer. Est-ce que cette propriété d'indépendance est valable pour n'importe quel langage rationnel ?

5.1.4 Monotonicité

Les fonctions f satisfaisant la condition $1 \leqslant f(n) \leqslant k^n$ ne sont pas toutes la complexité d'un mot infini. Caractériser ces fonctions est actuellement un problème ouvert. Cependant, des conditions nécessaires peuvent être données, et nous en verrons quelques-unes. Commençons par une condition simple.

Proposition 5.2. *La fonction de complexité est croissante : pour tout n, $p(n) \leqslant p(n+1)$.*

Preuve. Pour chaque facteur $w \in F_n(\mathbf{u})$, on peut choisir une occurrence $i(w)$ de w dans $\mathbf{u}$. Soit $e(w)$ le facteur de longueur $n+1$ qui apparaît à la position $i(w)$. Clairement, w est un préfixe de $e(w)$; en d'autres termes, $e(w)$ est un prolongement à droite de w. On définit ainsi une application e de $F_n(\mathbf{u})$ dans $F_{n+1}(\mathbf{u})$, qui est injective : deux mots différents de longueur n ne peuvent pas être préfixes du même mot de longueur $n+1$. Par conséquent le cardinal de $F_{n+1}(\mathbf{u})$ est supérieur ou égal à celui de $F_n(\mathbf{u})$: $p(n+1) \geqslant p(n)$. □

5.2 Le théorème de Morse et Hedlund

5.2.1 Énoncé et corollaires

Nous avons vu que les mots ultimement périodiques ont une complexité bornée. Le théorème suivant énonce que non seulement la réciproque est vraie, mais en fait la complexité d'un mot infini qui n'est pas ultimement périodique est très loin d'une fonction bornée.

Théorème 5.1 (Théorème de Morse et Hedlund)**.** *Soit $\mathbf{u}$ un mot infini et p sa fonction de complexité. Alors, soit $\mathbf{u}$ est ultimement périodique, soit p est strictement croissante.*

Corollaire 5.1. *S'il existe un entier n tel que $p(n) \leqslant n$, alors $\mathbf{u}$ est ultimement périodique.*

Preuve du corollaire 5.1. Si $\mathbf{u}$ n'est pas ultimement périodique, alors par le théorème 5.1, p est strictement croissante, i.e., $p(n+1) \geqslant p(n)+1$. Comme $p(0) = 1$, on en déduit que $p(n) \geqslant n+1$ pour tout n. □

Un autre corollaire est que des fonctions asymptotiquement équivalentes à $\sqrt{n}$ ou à $n/2$, par exemple, ne sont pas des fonctions de complexité.

5.2.2 Preuve du théorème

Preuve du théorème 5.1. Par la proposition 5.2, $p(n)$ est toujours croissante. Supposons qu'elle n'est pas strictement croissante ; alors il existe n tel que $p(n) = p(n+1)$. Soit $M = p(n)$.

Considérons l'application e définie dans la preuve de la proposition 5.2. C'est une injection entre deux ensembles finis de même cardinal, donc une bijection. Soit maintenant $f(w)$ le suffixe de longueur n de $e(w)$: f est une application (pas nécessairement injective) de $F_n(\mathbf{u})$ dans lui-même. Une application dans un ensemble fini a toujours un cycle : il existe x dans $F_n(\mathbf{u})$ et $m \leqslant M$ tels que $f^m(x) = x$ (prendre un y arbitraire et considérer les $M+1$ éléments y, $f(y)$, $f^2(y) = f(f(y))$, ..., $f^M(y)$; comme l'ensemble $F_n(\mathbf{u})$ a M éléments, deux d'entre eux doivent être égaux).

Supposons maintenant qu'un certain mot $w \in F_n(\mathbf{u})$ apparaît à la position i dans $\mathbf{u}$. Soit z le facteur de longueur $n+1$ qui apparaît à la même position i dans $\mathbf{u}$. Comme e est une bijection, $z = e(w')$ pour un certain $w' \in F_n(\mathbf{u})$; mais alors w' est un préfixe de z, de même que w, de sorte que $w' = w$ et $z = e(w)$. Le suffixe de longueur n de z est $f(w)$, et il apparaît à la position $i+1$ dans $\mathbf{u}$. En itérant cet argument, on conclut que $f^j(w)$ apparaît à la position $i+j$ dans $\mathbf{u}$.

En particulier, soit $w = x$. Alors $f^{j+m}(x) = f^j(x)$ apparaît aux positions $i+j+m$ et $i+j$ dans $\mathbf{u}$, de sorte que les lettres à ces positions sont les mêmes : $u_{i+j+m} = u_{i+j}$ pour tous $j \in \mathbb{N}$. On vient de montrer que $\mathbf{u}$ est ultimement m-périodique, avec une pré-période de longueur (au plus) i. □

L'ingrédient principal de la preuve ci-dessus est de discuter comment des facteurs peuvent être prolongés. Dans la section 5.4 nous développerons des outils pour cela, et avec eux la preuve deviendra plus courte.

Exercice 5.3. Soit $\mathbf{u}$ un mot infini qui n'est pas ultimement périodique. Soit $\ell(n)$ la différence maximale entre deux occurrences consécutives du même facteur de longueur n dans $\mathbf{u}$ (en supposant que cette différence est bornée : c'est le cas si $\mathbf{u}$ est uniformément récurrent). Montrer que $\ell(n) \geqslant n+1$ pour tout $n \in \mathbb{N}$.

5.2.3 Mots sturmiens

Le corollaire 5.1 est optimal : il existe des mots infinis de complexité égale à $n+1$ pour tout n. Un tel mot est appelé *mot sturmien*. La théorie des mots sturmiens est très développée, et nous n'entrerons pas dans le détail ici. Voir par exemple [7, chapitre 2].

Exemple 5.3. Le mot de Fibonacci

$$\varphi^\omega(a) = abaababaabaababaababaabaababaabaababaaba\dots$$

est un mot sturmien. Nous le prouverons plus tard, une fois que nous aurons introduit les outils appropriés dans la section 5.4.

Exercice 5.4. Soit $\mathbf{u}$ un mot sturmien. Supposons que aa apparaît dans $\mathbf{u}$, mais pas aaa. Soit $\mathbf{v}$ le mot infini obtenu en doublant chaque lettre de $\mathbf{u}$; par exemple, en partant du mot de Fibonacci on obtient

$$\mathbf{v} = aabbaaaabbaabbaaaabbaaaabbaabbaaaabbaabb\dots$$

Calculer la complexité de $\mathbf{v}$.

5.3 Complexité élevée

5.3.1 Sous-additivité

Proposition 5.3. *Soit p la fonction de complexité d'un mot infini $\mathbf{u}$. Alors, pour tous $m, n \in \mathbb{N}$, on a : $p(m+n) \leqslant p(m)p(n)$.*

En d'autre termes, la fonction à valeurs réelles $\log p$ est sous-additive.

Preuve. Pour $w \in F_{m+n}(\mathbf{u})$, soit w_1 son préfixe de longueur m et w_2 son suffixe de longueur n : alors $w = w_1 w_2$, avec $w_1 \in F_m(\mathbf{u})$ et $w_2 \in F_n(\mathbf{u})$. On définit ainsi une application injective de $F_{m+n}(\mathbf{u})$ dans $F_m(\mathbf{u}) \times F_n(\mathbf{u})$, donc $p(m+n) \leqslant p(m)p(n)$. □

Corollaire 5.2. *Si $F(\mathbf{u}) \neq A^*$, alors il existe $\alpha \in \mathbb{R}$ tel que $1 \leqslant \alpha < \#A$ et $p(n) = O(\alpha^n)$.*

Preuve. Soit $k = \#A$. Il existe m tel que $F_m(\mathbf{u}) \neq A^m$. Alors $p(m) < k^m$. Soit $\alpha = p(m)^{1/m}$: on a $1 \leqslant \alpha < k$.

Soit maintenant n un entier quelconque, et $q = \lceil n/m \rceil$, de sorte que $n \leqslant mq < m+n$. En itérant la proposition 5.3, on trouve que $p(mq) \leqslant p(m)^q$. Et par la proposition 5.2, $p(n) \leqslant p(mq)$. On a donc

$$p(n) \leqslant p(m)^q \leqslant \alpha^{mq} \leqslant \alpha^{m+n} = p(m)\alpha^n$$

et par conséquent $p(n) = O(\alpha^n)$. □

Une conséquence du corollaire 5.2 est qu'une fonction équivalente à $2^n/n$, par exemple, ne peut pas être la fonction de complexité d'un mot binaire.

Exercice 5.5. Montrer qu'une fonction équivalente à $2^n/n$ ne peut pas être la fonction de complexité d'un mot infini sur n'importe quel alphabet.

5.3.2 Entropie topologique

Proposition 5.4. *Soit p la fonction de complexité d'un mot infini $\mathbf{u}$. Alors la suite $\left(\frac{\log p(n)}{n}\right)_{n \geqslant 1}$ converge vers un réel h.*

Ce nombre h est appelé *entropie topologique* de $\mathbf{u}$.

Preuve. Soit $h = \inf_{n \geqslant 1} \frac{\log p(n)}{n}$. Nous allons montrer que $\frac{\log p(n)}{n}$ converge vers h. Fixons $\delta > 0$. Alors, par définition de la borne inférieure, il existe m tel que $\frac{\log p(m)}{m} < h + \delta$; soit $\alpha = p(m)^{1/m}$, de sorte que $\log \alpha < h + \delta$. On continue comme dans la preuve du corollaire 5.2 pour conclure que pour tout n, $p(n) \leqslant p(m)\alpha^n$. Alors $\frac{\log p(n)}{n} \leqslant \frac{\log p(m)}{n} + \log \alpha < \frac{\log p(m)}{n} + h + \delta$. Mais $\frac{\log p(m)}{n} < \delta$ pour n assez grand, et alors on a $h \leqslant \frac{\log p(n)}{n} < h + 2\delta$. Comme δ peut être choisi arbitrairement petit, ceci montre que la suite converge vers h. □

Exercice 5.6. Soit $\mathbf{u}$ un mot infini d'entropie nulle. Montrer que $F(\mathbf{u})$ est un langage rationnel si et seulement si $\mathbf{u}$ est ultimement périodique.

5.4 Outils pour la basse complexité

5.4.1 Facteurs spéciaux et bispéciaux

Dans la preuve du théorème 5.1, on a eu besoin de discuter comment un facteur peut être prolongé en un facteur plus long. Formalisons cela.

Soit $\mathbf{u} \in A^{\mathbb{N}}$ un mot infini, et w un facteur de $\mathbf{u}$. La *valence à droite* de w (dans $\mathbf{u}$) est le nombre de lettres différentes qui apparaissent dans $\mathbf{u}$ immédiatement après une occurrence de w :

$$d^+(w) = \#\{x \in A \colon wx \in F(\mathbf{u})\}.$$

La *valence à gauche* $d^-(w)$ est définie de façon similaire.

Un facteur de valence à droite au moins 2 est dit *spécial à droite* ; un facteur de valence à gauche au moins 2 est dit *spécial à gauche* ; et un facteur qui est à la fois spécial à droite et spécial à gauche est dit *bispécial*. On note respectivement $RS_n(\mathbf{u})$, $LS_n(\mathbf{u})$, et $BS_n(\mathbf{u})$ les ensembles des facteurs spéciaux à droite, spéciaux à gauche, et bispéciaux de longueur n dans $\mathbf{u}$.

Dans la preuve de la proposition 5.2, on a utilisé le fait que $d^+(w)$ est toujours au moins 1, de sorte que w est spécial à droite si et seulement si $d^+(w) \neq 1$. En revanche, $d^-(w)$ peut être nul, mais seulement dans des circonstances bien particulières : quand w est un préfixe de $\mathbf{u}$ qui n'a pas d'autre occurrence dans $\mathbf{u}$. Dans la plupart des exemples que nous rencontrerons, cela ne se produit pas car $\mathbf{u}$ est récurrent. Quand $\mathbf{u}$ est récurrent, les valences à droite et à gauche se comportent de façon symétrique.

Exercice 5.7. Soit $\ell(n)$ défini comme dans l'exercice 5.3. Montrer que s'il n'y a pas de facteur bispécial de longueur n, alors $\ell(n+1) = \ell(n+2)$.

5.4.2 Différences finies de la fonction de complexité

On définit maintenant de nouvelles fonctions qui dérivent de la fonction de complexité. Soit $s(n)$ la *première différence finie* de $p(n)$, i.e., $s(n) = p(n+1) - p(n)$; soit $b(n)$ sa *deuxième différence finie*, i.e., $b(n) = s(n+1) - s(n) = p(n+2) - 2p(n+1) + p(n)$. (La raison du choix des noms s et b apparaîtra plus tard).

Proposition 5.5. *Connaissant soit la fonction s, soit $p(1)$ et la fonction b, on peut retrouver la fonction p en utilisant les formules suivantes :*

$$p(n) = 1 + \sum_{l=0}^{n-1} s(l) = 1 + (p(1) - 1)n + \sum_{m=0}^{n-1} (n - 1 - m)b(m).$$

Preuve. La formule $p(n) = 1 + \sum_{l=0}^{n-1} s(l)$ est une conséquence immédiate de la définition de $s(n)$, comme $p(0) = 1$. De même, la définition de $b(n)$ donne $s(n) = s(0) + \sum_{m=0}^{n-1} b(m)$. En substituant la seconde formule

dans la première, sachant que $s(0) = p(1) - 1$, on obtient $p(n) = 1 + (p(1) - 1)n + \sum_{l=0}^{n-1} \sum_{m=0}^{l-1} b(m)$. Pour obtenir la formule désirée, il suffit d'inverser les deux sommes et de calculer $\sum_{l=m+1}^{n-1} b(m) = (n - 1 - m)b(m)$. □

La proposition 5.5 est particulièrement utile quand p croît lentement, car alors les fonctions s et b prennent en général de petites valeurs et sont plus faciles à évaluer. Par exemple, un mot infini est sturmien si et seulement si $s(n)$ vaut constamment 1, ou encore si et seulement si b est nulle et $p(1) = 2$.

5.4.3 Facteurs spéciaux et complexité

Pour évaluer s ou b, on peut utiliser les facteurs spéciaux ou bispéciaux.

Théorème 5.2. *Soit $\mathbf{u} \in A^*$ un mot infini, p sa fonction de complexité, et s et b ses deux premières différences finies. On a alors :*

(i) $s(n) = \sum_{w \in RS_n(\mathbf{u})} (d^+(w) - 1)$ est le nombre de facteurs spéciaux à droite de longueur n de $\mathbf{u}$, comptés avec multiplicité $d^+(w) - 1$.

(ii) Si $\mathbf{u}$ est récurrent, alors $s(n) = \sum_{w \in LS_n(\mathbf{u})} (d^-(w) - 1)$ est aussi le nombre de facteurs spéciaux à gauche de longueur n de $\mathbf{u}$, comptés avec multiplicité $d^-(w) - 1$.

(iii) Si $\mathbf{u}$ est récurrent, alors $b(n) = \sum_{w \in BS_n(\mathbf{u})} m(w)$ est le nombre de facteurs bispéciaux de longueur n de $\mathbf{u}$, comptés avec multiplicité

$$m(w) = \#\{(x, y) \in A^2 : xwy \in F(\mathbf{u})\} - d^+(w) - d^-(w) + 1.$$

Preuve. Comme les facteurs de longueur $n + 1$ peuvent être énumérés en énumérant d'abord leurs préfixes (resp. suffixes) de longueur n, puis pour chacun d'entre eux leurs prolongements à droite (resp. à gauche), on obtient :

$$p(n + 1) = \sum_{w \in F_n(\mathbf{u})} d^+(w) = \sum_{w \in F_n(\mathbf{u})} d^-(w).$$

Ainsi

$$s(n) = \sum_{w \in F_n(\mathbf{u})} (d^+(w) - 1) = \sum_{w \in F_n(\mathbf{u})} (d^-(w) - 1).$$

Mais $d^+(w) - 1$ est non nul si et seulement si w est spécial à droite ; et, si $\mathbf{u}$ est récurrent, $d^-(w) - 1$ est non nul si et seulement si w est spécial à gauche ; d'où (i) et (ii).

Pour obtenir (iii), observons d'abord que $m(w)$ peut être défini pour n'importe quel facteur w, mais est nul si w n'est pas bispécial. En sommant $m(w)$ sur $F_n(\mathbf{u})$, et en séparant les quatre termes, on obtient

$$\sum_{w \in F_n(\mathbf{u})} m(w) = p(n + 2) - p(n + 1) - p(n + 1) + p(n) = b(n).$$

□

Dans le cas d'un alphabet binaire, les valences des facteurs spéciaux sont toujours 2 et donc on peut se passer des multiplicités pour eux ; quant aux facteurs bispéciaux, leur multiplicité $m(w)$ ne peut valoir que -1, 0 ou 1. Un facteur bispécial est dit *fort* si $m(w) > 0$, *faible* si $m(w) < 0$, et *neutre* si $m(w) = 0$. Le théorème 5.2 devient alors :

Corollaire 5.3. *Soit $\mathbf{u}$ un mot infini binaire. Alors $s(n)$ est le nombre de ses facteurs spéciaux à droite de longueur n. Si $\mathbf{u}$ est récurrent, $s(n)$ est aussi le nombre de ses facteurs spéciaux à gauche de longueur n, et $b(n) = sb(n) - wb(n)$, où $sb(n)$ est le nombre de facteurs bispéciaux forts de longueur n, et $wb(n)$ le nombre de facteurs bispéciaux faibles de longueur n.*

Preuve. Soit $A = \{a, b\}$. Le seul argument non trivial est que $m(w) \in \{-1, 0, 1\}$. En effet, supposons que $\mathbf{u}$ est récurrent et w bispécial. Alors wa, wb, aw et bw sont tous facteurs de $\mathbf{u}$. Les mots aw et bw peuvent ne pas être spéciaux à droite, mais chacun a au moins un prolongement à droite, aws et bwt avec $s, t \in A$. L'ensemble $E = \{(x, y) \in A^2 : xwy \in F(\mathbf{u})\}$ contient au moins ces deux éléments, et au plus 4 éléments, de sorte que $2 \leqslant \#E \leqslant 4$, et on obtient $-1 \leqslant m(w) = \#E - 3 \leqslant 1$. □

La multiplicité $m(w)$ mesure d'une certaine manière la corrélation entre les prolongements à gauche et à droite de w. Dans le cas binaire, les facteurs bispéciaux faibles ont des prolongements complètement corrélés (une fois le prolongement à gauche choisi, celui à droite est imposé) tandis que les facteurs bispéciaux forts ont des prolongements non corrélés (les prolongements à gauche et à droite peuvent être choisis indépendamment).

5.4.4 Une autre preuve du théorème de Morse et Hedlund

En utilisant les facteurs spéciaux, on obtient une nouvelle preuve du théorème de Morse et Hedlund :

Preuve du théorème 5.1. Soit $\mathbf{u}$ un mot infini non ultimement périodique, et $n \in \mathbb{N}$. Comme $F_n(\mathbf{u})$ est fini, il existe au moins un facteur $w \in F_n(\mathbf{u})$ qui apparaît au moins deux fois, disons aux positions i et j, $i < j$. Il existe $m \in N$ tel que $u_{i+m} \neq u_{j+m}$, sinon un suffixe de $\mathbf{u}$ serait $j - i$-périodique ; et $m \geqslant n$ puisque le même mot w de longueur n apparaît aux positions i et j.

Soit $w' = u_i u_{i+1} \dots u_{i+m-1}$. Alors w' est spécial à droite, puisqu'il peut être prolongé aussi bien par u_{i+m} que par u_{j+m}, et il en est de même de son suffixe de longueur n. Il y a donc au moins un facteur spécial à droite de chaque longueur, et par le théorème 5.2, $s(n) = p(n+1) - p(n) \geqslant 1$. Il en résulte que p est strictement croissante. □

5.4.5 Graphes de Rauzy

Les prolongements des facteurs, leurs valences, les facteurs spéciaux et bispéciaux, etc. peuvent tous être visualisés en utilisant une représentation de $F_n(\mathbf{u})$ par un graphe, introduite par Gérard Rauzy en 1983.

Le *graphe de Rauzy* (or *graphe des facteurs*) d'ordre n de $\mathbf{u}$ est le graphe orienté G_n défini comme suit : G_n a $p(n)$ sommets étiquetés par les éléments de $F_n(\mathbf{u})$ (et identifiés à leurs étiquettes), et il y a une arête de v vers w si et seulement s'il existe deux lettres $x, y \in A$ telles que $vy = xw \in F_{n+1}(\mathbf{u})$. On décide d'étiqueter cette arête par la lettre x (parfois il est plus pratique de l'étiqueter par y, ou par le mot vy). Le graphe G_n a donc $p(n+1)$ arêtes.

Le degré entrant d'un sommet w est la valence à gauche $d^-(w)$, et son degré sortant est la valence à droite $d^+(w)$. Ainsi, un facteur est spécial à droite si et seulement si le sommet correspondant a au moins deux arêtes sortantes, et de même les facteurs spéciaux à gauche ou bispéciaux sont facilement reconnaissables.

Si on lit les facteurs de longueur n de $\mathbf{u}$ dans l'ordre de leurs occurrences, en commençant par $u_0 u_1 \dots u_{n-1}$, puis $u_1 u_2 \dots u_n$, etc., et on considère les sommets correspondants dans G_n, on obtient un chemin infini dans G_n, car deux sommets visités consécutivement sont toujours liés par une arête, et ce chemin infini est étiqueté par $\mathbf{u}$. De même, à partir d'un facteur z de longueur au moins n, on obtient un chemin fini dans G_n de longueur $|z| - n$, qui commence au sommet étiqueté par le préfixe de longueur n de z et se termine à celui étiqueté par son suffixe s de longueur n. Si l'étiquette de ce chemin est w, alors $z = ws$.

On peut voir G_n comme un automate fini non déterministe, dans lequel tous les états sont initiaux et finaux. Soit $L(G_n)$ le langage rationnel reconnu par G_n.

Proposition 5.6. *La suite de langages rationnels $L(G_n)$ approche $F(\mathbf{u})$ par dessus, i.e., $F(\mathbf{u}) = \bigcap_{n \in \mathbb{N}} L(G_n)$. Plus précisément, $L(G_{n+1}) \subset L(G_n)$, et si $m \leqslant n$, $L(G_n) \cap A^m = F_m(\mathbf{u})$.*

Preuve. Soit $w \in F(\mathbf{u})$. Alors w se prolonge en un mot $z = ws \in F_{|w|+n}(\mathbf{u})$, et z définit un chemin dans G_n étiqueté par w, si bien que $F(\mathbf{u}) \subset L(G_n)$.

Soit $w \in L(G_{n+1})$. Alors w est l'étiquette d'un chemin $(v_0, v_1, \dots v_{|w|})$ dans G_{n+1}. Soit v'_i le préfixe de longueur n de v_i : alors $(v'_0, v'_1, \dots, v'_{|w|})$ est un chemin dans G_n qui est aussi étiqueté par w, donc $L(G_{n+1}) \subset L(G_n)$.

Soit $m \leqslant n$ et $w \in L(G_n) \cap A^m$. Le mot w est l'étiquette d'un chemin de longueur m dans G_n, commençant à un certain sommet v. Le mot w est donc le préfixe de longueur m de v, donc $w \in F_m(\mathbf{u})$. Comme on sait déjà que $F_m(\mathbf{u}) \subset L(G_n)$, on obtient $L(G_n) \cap A^m = F_m(\mathbf{u})$.

Soit $L = \bigcap_{n \in \mathbb{N}} L(G_n)$. Pour tout m, $L \cap A^m = F_m(\mathbf{u})$, donc $L = F(\mathbf{u})$. □

Exercice 5.8. Montrer que $L(G_n) = L(G_{n+1})$ si et seulement si tout facteur bispécial w de longueur n, s'il en existe, a pour multiplicité $m(w) = (d^-(w) - 1)(d^+(w) - 1)$.

5.4.6 Application aux mots sturmiens

Lemme 5.1. *Les mots sturmiens sont récurrents.*

Preuve. Soit $\mathbf{u}$ un mot infini qui n'est pas récurrent. Il existe donc un mot w qui n'apparaît qu'un nombre fini de fois dans $\mathbf{u}$. Soit $n = |w|$ et $\mathbf{v}$ le mot obtenu en effaçant un préfixe suffisamment long de $\mathbf{u}$ de façon que $w \notin F(\mathbf{v})$. Alors $p_{\mathbf{v}}(n) < p_{\mathbf{u}}(n)$. Si $\mathbf{u}$ était sturmien, on aurait $p_{\mathbf{u}}(n) = n + 1$ de sorte que $p_{\mathbf{v}}(n) \leqslant n$, et $\mathbf{v}$ serait ultimement périodique par le corollaire 5.1. Mais alors $\mathbf{u}$ serait lui aussi ultimement périodique et $p_{\mathbf{u}}$ serait bornée, une contradiction. □

Si $\mathbf{u}$ est un mot sturmien, alors $s(n) = 1$, donc il y a exactement un facteur spécial à droite et un facteur spécial à gauche de chaque longueur. Il y a au plus un facteur bispécial de longueur n (si le facteur spécial à droite et le facteur spécial à gauche coïncident), et dans ce cas il est nécessairement neutre puisque $b(n) = 0$.

On peut donc décrire les graphes de Rauzy des mots sturmiens. S'il y a un facteur bispécial de longueur n, alors G_n consiste en deux boucles disjointes autour du sommet correspondant à ce facteur bispécial. Sinon, G_n a trois branches : une relie le facteur spécial à gauche au facteur spécial à droite, et les deux autres relient le facteur spécial à droite au facteur spécial à gauche.

5.5 Substitutions et complexité

Les substitutions sont l'un des outils les plus utiles pour construire et transformer des mots infinis non périodiques ayant des propriétés intéressantes en termes de complexité.

5.5.1 Un théorème de Cobham

Soit $\mathbf{u}$ un mot morphique. On peut prouver que le second morphisme peut toujours être supposé lettre-à-lettre :

Théorème 5.3 (Théorème de Cobham sur les mots morphiques)**.** *Un mot infini $\mathbf{u} \in A^{\mathbb{N}}$ est morphique si et seulement s'il peut s'écrire sous la forme $\mathbf{u} = g(f^\omega(x))$, où B est un alphabet, $x \in B$, $f\colon B^* \to B^*$ est un morphisme non effaçant, et $g\colon B^* \to A^*$ est un morphisme lettre-à-lettre.*

Exercice 5.9. Montrer que le mot caractéristique des puissances de 2, défini par $u_n = b$ s'il existe i tel que $n = 2^i$, $u_n = a$ sinon :

$$\mathbf{u} = abbabaaabaaaaaaabaaaaaaaaaaaaaab\dots$$

est morphique (indice : distinguer la première lettre des autres a) mais n'est pas purement morphique (indice : montrer qu'un mot purement morphique est récurrent si et seulement si sa première lettre apparaît au moins deux fois).

5.5.2 Le théorème d'Ehrenfeucht, Lee et Rozenberg

Théorème 5.4 ([5]). *La fonction de complexité d'un mot morphique vérifie $p(n) = O(n^2)$.*

Preuve. Soit $\mathbf{u}$ un mot morphique. Par le théorème 5.3, on peut l'écrire sous la forme $\mathbf{u} = g(\mathbf{v})$ avec $\mathbf{v}$ purement morphique et g lettre-à-lettre ; mais alors $F_n(\mathbf{u}) = g(F_n(\mathbf{v}))$, de sorte que $p_{\mathbf{u}}(n) \leqslant p_{\mathbf{v}}(n)$. Il suffit donc de prouver le théorème pour les mots purement morphiques engendrés par un morphisme non effaçant.

Supposons maintenant que $\mathbf{u} = f^\omega(a)$ avec $a \in A$ et $f\colon A^* \to A^*$ morphisme non effaçant et itérable à partir de a. Pour tout entier $N \geqslant 1$, on a aussi $\mathbf{u} = (f^N)^\omega(a)$. On peut donc remplacer f par l'un de ses itérés f^N de manière à garantir des propriétés supplémentaires.

Une lettre $x \in A$ est dite *bornée* si l'ensemble $\{f^i(x)\colon i \in \mathbb{N}\}$ est fini, et *croissante* sinon. Soit A_0 l'ensemble des lettres bornées et A_1 l'ensemble des lettres croissantes. Notons que $f(A_0) \subset A_0^+$ tandis que $f(A_1) \subset A^*A_1A^*$: l'image d'une lettre x contient une lettre croissante si et seulement si x est croissante. Comme la lettre a est nécessairement croissante, $A_1 \neq \emptyset$.

Soit E l'ensemble fini $\{f^i(x)\colon i \in \mathbb{N}, x \in A_0\}$. On définit une application ψ dans l'ensemble fini $A_1 \times E \times A_1$ comme suit : $\psi(x, w, y) = (x', f(w), y')$, où x' est la dernière lettre de A_1 qui apparaît dans $f(x)$, et y' est la première lettre de A_1 qui apparaît dans $f(y)$. L'itération de ψ conduit inévitablement à un cycle, donc il existe $N \geqslant 1$ tel que $\psi^{2N} = \psi^N$. En multipliant N si nécessaire, on peut aussi supposer que $|f^N(x)| \geqslant 2$ si x est une lettre croissante. En fait, il est suffisant de prendre $N = \#A!$.

En remplaçant f par f^N, on a les propriétés suivantes :

(i) Si $x \in A_0$, alors $f(x) \in A_0^+$ et $f(f(x)) = f(x)$.

(ii) Si $x \in A_1$, alors il existe deux lettres $y, z \in A_1$ telles que, pour tout $i \geqslant 1$, $f^i(x) \in A_0^* y A^* \cap A^* z A_0^*$.

(iii) Si $x \in A_1$, alors pour tout $i \in \mathbb{N}$, $|f^i(x)| \geqslant i + 1$.

Les propriétés (i) et (ii) viennent du fait que maintenant $\psi^2 = \psi$. La propriété (iii) est une conséquence du fait que $f^i(x)$ contient au moins une lettre croissante y, et que $|f(y)| \geqslant 2$, si bien que $|f^{i+1}(x)| > |f^i(x)|$.

Étudions maintenant $p(n)$. Fixons n, et considérons un facteur $w \in F_n(\mathbf{u})$. Le mot w est facteur de $f^k(x_0)$ pour un certain $k \geqslant 0$, que l'on supposera minimal, et $x_0 \in A$. Soit v un facteur de $f(x_0)$ de longueur minimale tel que w est facteur de $f^{k-1}(v)$, et v' un facteur de $f(v)$ de longueur minimale tel que w est facteur de $f^{k-2}(v')$. La minimalité de k implique que $|v| \geqslant 2$ et $|v'| \geqslant 2$.

Prouvons maintenant, par l'absurde, que $k < n+4$. Supposons que $k \geqslant n+4$, et discutons selon la nature des lettres de v'.

Premier cas : toutes les lettres de v' sont bornées, i.e., $v' \in A_0^*$. Alors $f^{k-2}(v') = f(v')$, et w est un facteur de $f(v')$ donc de $f^3(x_0)$, ce qui contredit la minimalité de k.

Deuxième cas : v' contient une lettre croissante interne, i.e., $v' = v_1 x v_2$ avec $v_1 \in A^+$, $x \in A_1$, et $v_2 \in A^+$. Alors $f^{k-2}(x)$ est facteur de w, sinon $|v'|$ ne serait pas minimal. Mais $|f^{k-2}(x)| \geqslant k - 1 > n$, une contradiction.

Troisième cas : seule la dernière lettre de v' est croissante, i.e., $v' = ry$ avec $y \in A_1$ et $r \in A_0^+$ (on rappelle que $|v'| \geqslant 2$, donc $r \neq \varepsilon$). Soit x la dernière lettre de v ; alors un suffixe $r'y$ de v' est préfixe de $f(x)$, sinon $|v|$ ne serait pas minimal. La propriété (ii) implique que $f(y) = syt$ avec $s \in A_0^*$. Comme v' est minimal, on peut écrire $w = w_1 w_2$ où w_1 est un suffixe de $f^{k-2}(r) = f(r) = f^{k-3}(r)$ et w_2 est un préfixe de $f^{k-2}(y)$. Si $s \neq \varepsilon$, alors $f(s)^{k-3}$ est un préfixe de $f^{k-2}(y)$, et, comme $|f(s)^{k-4}| \geqslant n$, w_2 est un préfixe de $f(s)^{k-4}$, et donc aussi de $f^{k-3}(y)$. Si $s = \varepsilon$, alors $f^{k-2}(y) = f^{k-3}(yt)$, et, comme $|f^{k-3}(y)| \geqslant k - 2 > n$, w_2 est aussi préfixe de $f^{k-3}(y)$. Dans les deux sous-cas, on trouve que w apparaît déjà dans $f^{k-3}(v')$, en contradiction avec la minimalité de k.

Quatrième cas : seule la première lettre de v' est croissante. Ce cas est symétrique du troisième.

Cinquième cas : seules la première et la dernière lettres de v' sont croissantes. Alors les arguments des troisième et quatrième cas peuvent être appliqués ensemble pour arriver à nouveau à une contradiction.

Soit $v = v_1 v_2$ avec $|v_1| = 1$ et $|v_2| \geqslant 1$. Comme v est minimal, on peut écrire $w = w_1 w_2$ avec w_1 suffixe de $f^{k-1}(v_1)$ et w_2 préfixe de $f^{k-1}(v_2)$. Soit $i = |w_1|$. Le mot w est complètement déterminé par la donnée de v, k, et i. Comme $|v|$ est borné, $k < n + 4$ et $i < n$, il y a $O(n^2)$ manières de choisir ces données. □

5.5.3 Le théorème de Pansiot

Dans le cas des mots purement morphiques, Le théorème 5.4 peut être affiné pour obtenir des estimations precises de la complexité. On rappelle que $f(n) = \Theta(g(n))$ signifie qu'il existe deux constantes réelles strictement positives C_1 et C_2 telles que, pour tout n assez grand, $C_1 g(n) \leqslant f(n) \leqslant C_2 g(n)$.

Théorème 5.5 ([10]). *Soit* $\mathbf{u}$ *un mot purement morphique et* p *sa fonction de complexité. Alors l'une des propriétés suivantes est satisfaite :*
(i) $p(n) = \Theta(1)$ *;*
(ii) $p(n) = \Theta(n)$ *;*
(iii) $p(n) = \Theta(n \log\log n)$ *;*
(iv) $p(n) = \Theta(n \log n)$ *;*
(v) $p(n) = \Theta(n^2)$.

5.6 Exemples de calculs de complexité

Intéressons nous maintenant au problème pratique de trouver une expression de $p(n)$ pour un mot infini donné. Nous allons étudier quelques exemples. D'autres exembles peuvent être trouvés par exemple dans [1] ou dans [4].

5.6.1 Le mot caractéristique des puissances de 2

Soit $\mathbf{u}$ défini par $u_n = b$ s'il existe i tel que $n = 2^i$, $u_n = a$ sinon :

$$\mathbf{u} = abbabaaabaaaaaaabaaaaaaaaaaaaaab\ldots$$

Ce mot n'est pas récurrent. Tout facteur qui contient au moins deux occurrences de b n'apparaît qu'une fois, car la distance entre deux occurrences consécutives de b ne prend jamais la même valeur.

Pour calculer la complexité, nous allons compter les facteurs en fonction du nombre de b qu'ils contiennent. Fixons $n \in \mathbb{N}$.

Premier cas : facteurs qui ne contiennent pas de b. Il y a un seul mot de chaque longueur qui ne contient pas de b, a^n, et il est facteur de $\mathbf{u}$.

Deuxième cas : facteurs qui contiennent exactement un b. Il y a n mots de longueur n qui contiennent exactement un b (caractérisés par la position du b), et tous sont facteurs de $\mathbf{u}$.

Troisième cas : facteurs qui contiennent au moins deux occurrences de b. De telles facteurs apparaissent une seul fois, donc il suffit de compter les positions où ils apparaissent. Soit w le facteur de longueur n qui commence à la position j.

Premier sous-cas : $n \leqslant 2$. La seule possibilité est $j = 2$, $w = bb$.

Deuxième sous-cas : $3.2^{i-1} \leqslant n \leqslant 2^{i+1}$ pour un certain $i \geqslant 1$. Si $0 \leqslant j \leqslant 2^{i-1}$, alors w contient un b aux positions $2^{i-1} - j$ et $2^i - j$. Si $2^{i-1} + 1 \leqslant j \leqslant 2^i$, alors w contient un b aux positions $2^i - j$ et $2^{i+1} - j$. Si $j > 2^i$, alors w contient au plus un b, car $n < 2^{i+2} - 2^{i+1} + 1$. En tout nous avons $2^i + 1$ possibilités.

Troisième sous-cas : $2^{i+1} + 1 \leqslant n \leqslant 3.2^i - 1$ pour un certain $i \geqslant 1$. Si $0 \leqslant j \leqslant 2^i$, alors w contient au moins deux occurrences de b comme ci-dessus. Si $2^i + 1 \leqslant j \leqslant 2^{i+2} - n$, alors w contient un seul b, à la position $2^{i+1} - j$. Si $2^{i+2} + 1 - n \leqslant j \leqslant 2^{i+1}$, alors w contient un b aux positions $2^{i+1} - j$ et $2^{i+2} - j$. Si $j > 2^{i+1}$, alors w contient au plus un b, car $n < 2^{i+3} - 2^{i+2} + 1$. En tout nous avons $n - 2^i + 1$ possibilités.

En tenant compte des différents cas, nous obtenons :
– p(0)=1, p(1)=2, p(2)=4 ;
– si $3.2^{i-1} \leqslant n \leqslant 2^{i+1}$ pour un certain $i \geqslant 1$, alors $p(n) = n + 2^i + 2$;
– si $2^{i+1} \leqslant n \leqslant 3.2^i$ pour un certain $i \geqslant 1$, alors $p(n) = 2n - 2^i + 2$.

Nous aurions pu obtenir le même résultat avec moins de calculs en comptant seulement les facteurs spéciaux à droite. En effet, un facteur spécial apparaît au moins deux fois, donc il ne peut pas contenir deux occurrences de b. Seuls les deux premiers cas subsistent.

Premier cas : facteurs qui ne contiennent pas de b. Le facteur a^n est spécial à droite.

Deuxième cas : facteurs qui contiennent exactement un b. Supposons que $w = a^k b a^j$ est spécial à droite. Alors $wb = a^k b a^j b$ apparaît, donc $j = 2^i - 1$ pour un certain i. Si $i = 0$, on trouve que $w = b$ ou $w = ab$. Sinon, wb doit être un facteur de $ba^{2^{i-1}-1}ba^{2^i-1}b$, donc $k \leqslant 2^{i-1} - 1$. Comme $k + 1 + j = n$, cela se produit si et seulement si $0 \leqslant n - 2^i \leqslant 2^{i-1} - 1$, i.e., $2^i \leqslant n \leqslant 3.2^{i-1} - 1$.

On trouve donc que $s(n) = 1$ si $n = 0$ ou $3.2^{i-1} \leqslant n \leqslant 2^{i+1} - 1$ pour un certain $i \geqslant 1$, $s(n) = 2$ si $n = 1$ ou $2^{i+1} \leqslant n \leqslant 3.2^i - 1$ pour un certain $i \geqslant 1$, et $s(2) = 3$. Alors la proposition 5.5 permet de calculer $p(n)$.

5.6.2 Le mot de Fibonacci

Rappelons que le mot de Fibonacci est $\mathbf{u} = \varphi^\omega(a)$ où $\varphi(a) = ab$ et $\varphi(b) = a$. L'image miroir d'un mot $w = a_1 a_2 \dots a_n$ est le mot $\tilde{w} = a_n a_{n-1} \dots a_1$.

Lemme 5.2. *Le langage du mot de Fibonacci word est fermé par image miroir.*

Preuve. Soit $\tilde{\varphi}$ le morphisme défini par $\tilde{\varphi}(a) = ba$ et $\tilde{\varphi}(b) = a$. Observons d'abord que, pour tout $w \in \{a, b\}^*$, l'image miroir de $\varphi(w)$ est $\widetilde{\varphi(w)} = \tilde{\varphi}(\tilde{w})$. De plus, pour tout $w \in \{a, b\}^*$, $a\tilde{\varphi}(w) = \varphi(w)a$.

Montrons maintenant par récurrence sur i que l'image mirroir de $f_i = \varphi^i(a)$ est un facteur de $\mathbf{u}$. C'est trivialement vrai pour $i = 0$. Supposons que $\tilde{f}_i \in F(\mathbf{u})$. Alors $\tilde{f}_{i+1} = \widetilde{\varphi(f_i)} = \tilde{\varphi}(\tilde{f}_i)$. Nous savons que $\tilde{f}_i$ est facteur de $\mathbf{u}$; ce facteur se prolonge en un facteur $\tilde{f}_i x$, avec $x \in \{a, b\}$. Alors $\varphi(\tilde{f}_i x)$ est aussi facteur de $\mathbf{u}$, et donc son préfixe $\varphi(\tilde{f}_i)a = a\tilde{\varphi}(\tilde{f}_i) = a\tilde{f}_{i+1}$ est facteur de $\mathbf{u}$, donc $\tilde{f}_{i+1}$ est facteur de $\mathbf{u}$.

Comme tout facteur de $\mathbf{u}$ est facteur d'un certain f_i, on conclut que $F(\mathbf{u})$ est fermé par image miroir. □

Lemme 5.3. *Si $\varphi(w)a$ est un facteur de $\mathbf{u}$, alors w est un facteur de $\mathbf{u}$.*

Preuve. Comme $\mathbf{u} = \varphi(\mathbf{u})$, il existe $v \in F(\mathbf{u})$ tel que $\varphi(w)a$ est facteur de $\varphi(v)$, et on peut choisir v de longueur minimale. Alors $\varphi(w)a = t_1\varphi(v')t_2$, où $v = x_1 v' x_2$, $t_1 \in \{a, b, ab\}$ est un suffixe non vide de $\varphi(x_1)$, et $t_2 \in \{a, ab\}$ est un préfixe non vide de $\varphi(x_2)$. Mais $t_1 = b$ est impossible, car $\varphi(w)a$ commence par un a. Donc t_1 est soit a soit ab, et dans les deux cas $t_1 = \varphi(x_1)$. De même, t_2 ne peut être ab, car $\varphi(w)a$ se termine par un a, donc $t_2 = a$. Nous avons alors $\varphi(w)a = \varphi(x_1 v')a$, et comme $\{\varphi(a), \varphi(b)\}$ est un code, φ est injectif et nous concluons que $w = x_1 v'$, qui est donc facteur de $\mathbf{u}$. □

Proposition 5.7. *Les facteurs spéciaux à gauche du mot de Fibonacci sont ses préfixes. Ses facteurs spéciaux à droite sont les images mirroir de ses préfixes.*

Preuve. Si w est spécial à gauche, alors $\varphi(w)$ est aussi spécial à gauche. En effet, si aw et bw sont tous deux facteurs de $\mathbf{u}$, alors il en est de même de $ab\varphi(w)$ et $a\varphi(w)$. En itérant cette propriété, on obtient que tous les $f_i = \varphi^i(a)$ sont spéciaux à gauche, et donc tous les préfixes de $\mathbf{u}$ sont spéciaux à gauche (puisqu'un préfixe d'un facteur spécial à gauche est spécial à gauche).

Réciproquement, supposons que w est spécial à gauche. Nécessairement w commence par un a. Si w se termine par un b, il est toujours suivi par un a donc $w' = wa$ est aussi spécial à gauche. Sinon on pose $w' = w$. Alors w' se termine par a et ne contient pas bb, donc il se factorise sur $\{ab, a\}$, le dernier facteur étant a. Il existe alors un mot v tel que $w' = \varphi(v)a$. Par le lemme 5.3, cela implique que v est un facteur de $\mathbf{u}$. De plus, on sait que w' est spécial à gauche, donc aw' et bw' sont facteurs de $\mathbf{u}$, et ce dernier est toujours précédé de a. On peut appliquer le lemme 5.3 à nouveau à $aw' = \varphi(bv)a$ et à $abw' = \varphi(av)a$ pour conclure que bv et av sont facteurs de $\mathbf{u}$, i.e., v est spécial à gauche. Si $v \neq \varepsilon$, on a $|v| < |\varphi(v)|$ (puisque v commence par un a) et $|\varphi(v)| = |w'| - 1 \leqslant |w|$. On peut donc procéder par récurrence : supposons que v est un préfixe de $\mathbf{u}$. Alors vx est un préfixe de $\mathbf{u}$ pour une certaine lettre x, et alors w est un préfixe de $w' = \varphi(v)a$ qui est un préfixe de $\varphi(vx)$ qui est un préfixe de $\mathbf{u}$.

Le lemme 5.2 implique que les facteurs spéciaux à droite sont les images mirroir des facteurs spéciaux à gauche. □

Corollaire 5.4. *Le mot de Fibonacci est sturmien, i.e., sa complexité est $p(n) = n + 1$.*

Preuve. La proposition 5.7 implique qu'il y a exactement un facteur spécial à droite de chaque longueur. Par le corollaire 5.3, on conclut que $s(n) = 1$ pour tout n et donc $p(n) = n + 1$. □

Le principal argument dans la preuve est le lemme 5.3. Il permet de "désubstituer" φ, et d'associer à chaque facteur spécial un facteur spécial plus court, et de remonter ainsi jusqu'au mot vide. Un argument de ce type est toujours utilisé pour calculer la complexité d'un mot morphique.

5.6.3 Le mot de Thue-Morse

Rappelons que le mot de Thue-Morse est le mot purement morphique

$$\mathbf{t} = abbabaabbaababbabaababbaabbabaabbaababbaabbabaababbabaabbaab\dots$$

point fixe commençant par a du morphisme θ défini par $\theta(a) = ab$ et $\theta(b) = ba$.

Comme il est récurrent, on peut utiliser les facteurs bispéciaux.

Proposition 5.8. *Les facteurs bispéciaux de* $\mathbf{t}$ *sont*
- *facteurs bispéciaux forts :* ε, $\theta^m(ab)$, $\theta^m(ba)$ *pour* $m \geqslant 0$;
- *facteurs bispéciaux neutres :* a, b ;
- *facteurs bispéciaux faibles :* $\theta^m(aba)$, $\theta^m(bab)$ *pour* $m \geqslant 0$.

La preuve repose sur le lemme suivant, qui joue le même rôle que le lemme 5.3 pour le mot de Fibonacci :

Lemme 5.4. *Tout facteur de* $\mathbf{t}$ *est de la forme* $w = r_1.\theta(v).r_2$ *avec* $v \in F(\mathbf{t})$ *et* $r_i \in \{\varepsilon, a, b\}$. *Si* $|w| \geqslant 5$ *cette décomposition est unique.*

Preuve de la proposition 5.8. Soit w un facteur tel que $|w| \geqslant 5$. D'après le lemme, w s'écrit de manière unique $w = r_1.\theta(v).r_2$. Supposons que $r_2 = a$. Si le mot wa était facteur de $\mathbf{t}$, on pourrait aussi lui appliquer le lemme : $wa = r_1'.\theta(v').r_2'$. Comme $\theta(v')$ ne peut pas se terminer par aa, on aurait nécessairement $r_2' = a$, donc $w = r_1'.\theta(v').\varepsilon$, ce qui contredirait l'unicité dans le lemme. Le facteur w ne peut donc se prolonger à droite que par un b. De même, si $r_2 = b$ le seul prolongement possible est wa. Si w est spécial à droite, alors nécessairement $r_2 = \varepsilon$. Le prolongement wa se décompose alors sous la forme $wa = r_1.\theta(v).a$, donc a un prolongement unique wab dont la décomposition est nécessairement $r_1.\theta(va).\varepsilon$, donc va est facteur de $\mathbf{t}$, et pour la même raison vb est aussi facteur de $\mathbf{t}$, donc v est spécial à droite.

On voit donc que les facteurs spéciaux à droite de longueur au moins 5 sont tous de la forme $w = r_1\theta(v)$ avec v lui-même spécial à droite. Une propriété symétrique est vérifiée à gauche, donc les facteurs bispéciaux de longueur au moins 5 sont tous des images par θ de facteurs bispéciaux, qui de plus sont de même type.

Il ne reste plus qu'à énumérer les facteurs bispéciaux courts pour pouvoir conclure : ε, ab, ba, $abba$, $baab$ sont forts, a et b sont neutres, et aba et bab sont faibles. □

On en déduit des formules pour $s(n)$ et $p(n)$.

Corollaire 5.5. *Pour le mot de Thue-Morse, on a*

$$s(n) = \begin{cases} 1 & si\ n = 0 \\ 2 & si\ 0 < n \leqslant 2 \\ 4 & si\ 2.2^m < n \leqslant 3.2^m \\ 2 & si\ 3.2^m < n \leqslant 4.2^m \end{cases} \quad et$$

$$p(n) = \begin{cases} 1 & si\ n = 0 \\ 2 & si\ n = 1 \\ 4 & si\ n = 2 \\ 4n - 2.2^m - 4 & si\ 2.2^m < n \leqslant 3.2^m \\ 2n + 4.2^m - 2 & si\ 3.2^m < n \leqslant 4.2^m \end{cases},$$

où $m \in \mathbb{N}$.

Exercice 5.10. Appliquer la même méthode au *mot de doublement de période*, point fixe du morphisme f défini par $f(a) = ab$ et $f(b) = aa$.

5.6.4 Un example de complexité légèrement supérieure

Soit $\mathbf{u} = f^\omega(a)$ avec $f(a) = aba$ et $f(b) = bb$:

$$\mathbf{u} = ababbabababbbabababbabbbbbbbbbababbabababbbabababbabbbbbbbbbb\ldots$$

Ici le lemme clé est

Lemme 5.5. *Si $w \in F(\mathbf{u})$, alors $w = r_1.f(v).r_2$ avec $r_1 \in \{\varepsilon, a, b, ba\}$, $v \in F(\mathbf{u})$ et $r_2 \in \{\varepsilon, a, b, ab\}$. Si de plus $|w| \geqslant 4$ et $w \notin b^*$, alors cette décomposition est unique.*

Le seul facteur bispécial de longueur inférieure à 4 qui n'est pas une puissnce de b est bab, et il engendre une famille $f^m(bab)$ de facteurs bispéciaux faibles, de longueur $2^m(3+m/2)$. Les puissances de b sont toutes des facteurs bispéciaux, forts si la longueur est une puissance de 2, neutres sinon.

On a : $sb(n) = 1$ si $n = 2^m$, 0 sinon ; $wb(n) = 1$ si $n = 2^m(3+m/2)$, 0 sinon. On en déduit $s(n) = 1+q-r$, où $q = \lceil \log_2 n \rceil$ est le nombre de facteurs bispéciaux forts de longueur inférieure à n, et $r = \left\lceil \frac{W(128n\log 2)}{\log 2} - 6 \right\rceil$ est le nombre de facteurs bispéciaux faibles de longueur inférieure à n, W étant l'unique fonction analytique réelle sur $]-1/e, +\infty[$ telle que $W(x)e^{W(x)} = x$.

Finalement,

$$\begin{aligned} p(n) &= 1 + \sum_{m=0}^{n-1} s(m) \\ &= 1 + n + \sum_{m=0}^{q-1} (n - 1 - 2^m) - \sum_{m=0}^{r-1} (n - 1 - 2^m(3 + m/2)) \\ &= 1 + n + (n-1)q - (2^q - 1) - (n-1)r + (2^{r+1} + 2^{r-1}r - 2) \ . \end{aligned}$$

Dans cette expression, les termes dominants sont nq et nr, soit

$$p(n) = n\frac{\log n - W(128n\log 2)}{\log 2} + O(n) \ .$$

On peut alors, en utilisant le fait que $W(x) = \log x - \log\log x + O\left(\frac{\log\log x}{\log x}\right)$ obtenir un équivalent de la complexité de la suite étudiée : $p(n) \sim n \log_2 \log_2 n$.

Références

[1] J.-P. Allouche, Sur la complexité des suites infinies, *Bull. Belg. Math. Soc.* **1** (1994), 133–143.

[2] J.-P. Allouche et J. O. Shallit, *Automatic Sequences, Theory, Applications, Generalizations*, Cambridge University Press, 2003.

[3] J. Berstel et J. Karhumäki, Combinatorics on words : a tutorial, *Bull. European Assoc. Theor. Comput. Sci.* **79** (2003), 178–228.

[4] V. Berthé et M. Rigo (éd.), *Combinatorics, Automata and Number Theory*, *Encyclopedia of Mathematics and its Applications* **135**, Cambridge University Press, 2010.

[5] A. Ehrenfeucht, K. P. Lee et G. Rozenberg, Subword complexities of various classes of deterministic developmental languages without interaction, *Theoret. Comput. Sci.* **1** (1975), 59–75.

[6] M. Lothaire, *Combinatorics on Words*, *Encyclopedia of Mathematics and its Applications* **17**, Addison-Wesley, 1983.

[7] M. Lothaire, *Algebraic Combinatorics on Words*, *Encyclopedia of Mathematics and its Applications* **90**, Cambridge University Press, 2002.

[8] M. Lothaire, *Applied Combinatorics on Words*, *Encyclopedia of Mathematics and its Applications* **105**, Cambridge University Press, 2005.

[9] M. Morse et G. A. Hedlund, Symbolic dynamics, *Amer. J. Math.* **60** (1938), 815–866.

[10] J.-J. Pansiot, Complexité des facteurs des mots infinis engendrés par morphismes itérés, in *ICALP '84*, pp. 380–389, *Lect. Notes Comp. Sci.* **172**, Springer-Verlag, 1984.

[11] N. Pytheas Fogg, *Substitutions in Dynamics, Arithmetics and Combinatorics*, *Lecture Notes in Mathematics* **1794**, Springer Verlag, 2002. Ed. by V. Berthé, S. Ferenczi, C. Mauduit, and A. Siegel.