Jouer avec les mots, pourquoi et comment ?

Michel Rigo*

Résumé

Ce texte reprend l'essentiel de ma présentation à la *Brussels Math. Summer School* du 4 août 2015. Il s'agit d'une courte introduction à la *combinatoire des mots*. A l'instar de Raymond Queneau et ses cent mille milliards de poèmes, nous construisons des suites aux propriétés surprenantes. Pour ne pas allonger le texte, nous avons décidé d'éviter l'emploi d'automates finis.

Les premiers résultats en combinatoire des mots remontent au début du siècle précédent, avec les travaux du mathématicien norvégien Axel Thue. Cette branche des mathématiques étudie la structure et les arrangements apparaissant au sein de suites finies, ou infinies, de symboles appartenant à un ensemble fini.

Un carré est la juxtaposition de deux répétitions d'un même mot. On dira qu'un mot comme "taratata" contient un carré. Il est aisé de vérifier que, si on dispose uniquement de deux symboles, alors tout mot de longueur au moins 4 contient un carré. Cette observation amène de nombreuses questions simples à formuler : Avec trois symboles, peut-on construire un mot arbitrairement long ne contenant pas de carré ? Si on se limite à deux symboles, peut-on construire un mot arbitrairement long sans cube, i.e., évitant la juxtaposition de trois répétitions d'un même mot ? En fonction de la taille de l'alphabet, quels motifs doivent nécessairement apparaître et quels sont ceux qui sont évitables ? Que se passe-t-il si on autorise certaines permutations ?

*Université de Liège, M.Rigo@ulg.ac.be

1 Exemple introductif

Dans cette première section, nous partons d'un simple exercice pour motiver l'introduction des notions de *mots* (finis et infinis) et de *morphismes* qui seront définies rigoureusement dans les sections suivantes.

Soit $n \geq 0$ un entier. Commençons par un exercice d'échauffement, à savoir l'étude du signe, sur l'intervalle $[0,\pi]$, de la fonction

$$F_n(x) := \sin(x)\sin(2x)\sin(4x)\cdots\sin(2^n x)$$

obtenue comme produit de n facteurs où l'on double, à chaque fois, l'argument. Pour $n = 0$, la fonction $F_0(x) = \sin(x)$ est trivialement positive [1] sur l'intervalle considéré. Pour $n = 1$, tout comme $\sin(2x)$, la fonction $F_1(x)$ est positive sur $[0,\pi/2]$ puis négative sur $[\pi/2,\pi]$. La figure 0.1 reprend les quatre fonctions à multiplier pour déterminer le signe de F_3.

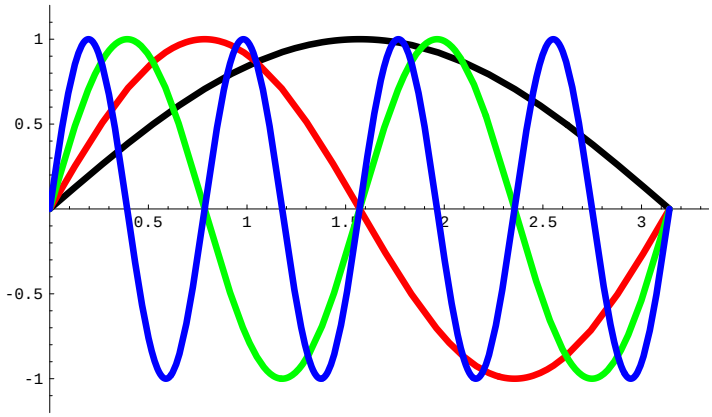


FIGURE 0.1 — Les quatre fonctions définissant F_3.

Pour n quelconque, on s'aperçoit rapidement qu'il convient de diviser l'intervalle $[0,\pi]$ en 2^n sous-intervalles

$$I_{n,j} := \left[\frac{j\pi}{2^n}, \frac{(j+1)\pi}{2^n}\right] \quad j = 0,\ldots,2^n-1$$

puisque la fonction $\sin(2^n x)$ change $2^n - 1$ fois de signe.

Remarque 1. L'idée de ce petit exercice introductif est due à Jean-Paul Allouche, voir [2]. Contrairement à la duplication des sinus, une 'formule directe' existe pour la duplication des cosinus

$$\cos(x)\cos(2x)\cos(4x)\cdots\cos(2^n x) = \frac{\sin(2^{n+1}x)}{2^{n+1}\sin(x)}.$$

1. On doit comprendre *positive ou nulle*, mais les points où la fonction F_n est nulle sont facilement caractérisés par les bornes $j\pi/2^n$ des intervalles $I_{n,j}$ et ne nous intéresseront donc pas.

Dans la table 0.1, nous avons retranscrit, pour les premières valeurs de n, la suite des 2^n signes pris par F_n sur les sous-intervalles successifs $I_{n,0}, I_{n,1}, \ldots, I_{n,2^n-1}$. De façon symbolique, on notera simplement + ou − selon le cas.

F_0	+															
F_1	+	-														
F_2	+	-	-	+												
F_3	+	-	-	+	-	+	+	-								
F_4	+	-	-	+	-	+	+	-	-	+	+	-	+	-	-	+

TABLE 0.1 — Signes de F_n sur la subdivision *ad hoc* de $[0, \pi]$, $n = 0, 1, 2, 3, 4$.

On peut dès lors poser la question suivante : *est-il possible de caractériser 'facilement' le signe de F_n sur l'intervalle $I_{n,j}$, et ce, quel que soit n?* Bien sûr, par facilement, il faut entendre à la suite d'un calcul (ou d'un algorithme) dont le nombre d'opérations (la complexité) soit "petit" par rapport à n, par exemple, que ce nombre soit proportionnel à $\log n$. En guise d'illustration, pourriez-vous déterminer le signe de F_{30} dans l'intervalle $I_{30,17\,893\,617}$?

Remarque 2. Si on dispose de la suite des 2^n signes pris par F_n, alors on trouve aisément la suite des 2^{n+1} signes pris par F_{n+1}. En effet, puisque

$$F_{n+1}(x) = F_n(x) \sin(2^{n+1} x)$$

et que, sur l'intervalle $I_{n,j}$, la fonction $\sin(2^{n+1} x)$ est d'abord positive puis négative, on en déduit que si F_n est positive (resp. négative) sur $I_{n,j}$, alors F_{n+1} est positive (resp. négative) sur $I_{n+1,2j}$ puis négative (resp. positive) sur $I_{n+1,2j+1}$.

A ce stade, cette remarque nous permet uniquement de déterminer la suite de 2^{n+1} signes associés à F_{n+1} à partir de la suite de 2^n signes associés à F_n. Il faut donc conserver une quantité d'information très importante !

2 Un brin de formalisme

Commençons par présenter quelques définitions élémentaires (mots finis et infinis). Comme ouvrages de référence en combinatoire des mots, citons [16, 18].

Définition 3. Un *alphabet* est un ensemble fini. Ainsi,

$$\{a, b, c\},\ \{\heartsuit, \diamond, \clubsuit, \spadesuit\},\ \{0, 1\}, \{\rightarrow, \leftarrow, \uparrow, \downarrow\}$$

ou encore $\{+, -\}$ sont des alphabets. Les éléments d'un alphabet sont appelés *lettres* ou *symboles*.

Définition 4. Soit A un alphabet. Un *mot* (fini) sur A est une suite finie (et ordonnée) de symboles, i.e., une application de $\{0, \ldots, n-1\}$ dans A. Par exemple, *abbac* et *ba* sont deux mots sur l'alphabet $\{a, b, c\}$. La *longueur* d'un mot w est le nombre de symboles constituant ce mot ; on la note $|w|$. Ainsi,

$$|abbac| = 5 \ \text{ et } \ |ba| = 2.$$

L'unique mot de longueur 0 est le mot correspondant à la suite vide. Ce mot s'appelle le *mot vide* et on le note ε. L'ensemble des mots sur A est noté A^*. Par exemple,

$$\{a,b,c\}^* = \{\varepsilon, a, b, c, aa, ab, ac, ba, bb, bc, ca, cb, cc, aaa, aab, \dots\}.$$

Remarque 5. L'ensemble A^* des mots finis sur A muni de l'opération de *concaténation* est un monoïde [2] dont le neutre est ε. Sans donner de définition formelle, voici un exemple de concaténation :

$$bon \cdot jour = bonjour.$$

On remarque que $|u \cdot v| = |u| + |v|$ pour tous $u, v \in A^*$.

Définition 6. Soit A un alphabet. Un *mot infini* sur A est simplement une application $\mathbf{w} : \mathbb{N} \to A$, i.e., une suite d'éléments de A. Ainsi, de façon classique, on notera $A^{\mathbb{N}}$ l'ensemble des mots infinis sur A. Pour rappel, A^B dénote l'ensemble des applications de B dans A. Par exemple, voici le préfixe d'un mot infini sur l'alphabet $\{0,\dots,9\}$ (le développement décimal de $\pi - 3$) :

$$14159265358979323846264338327950288419716939937\cdots.$$

Dans ces notes, pour distinguer les mots infinis des mots finis, nous utiliserons des lettres en gras.

Nous voudrions à présent définir la notion de suite de mots infinis convergeant vers un mot infini limite, puis une suite de mots *finis* convergeant vers un mot infini limite. Pour ce faire, nous munissons l'ensemble $A^{\mathbb{N}}$ d'une distance

$$d : A^{\mathbb{N}} \times A^{\mathbb{N}} \to [0, +\infty[$$

définie comme suit. Si $\mathbf{x}$ et $\mathbf{y}$ sont deux mots infinis, alors $\mathbf{x} \wedge \mathbf{y}$ désigne leur plus long préfixe commun. Si $\mathbf{x} = \mathbf{y}$, alors on pose $d(\mathbf{x}, \mathbf{y}) = 0$, sinon

$$d(\mathbf{x}, \mathbf{y}) = 2^{-|\mathbf{x} \wedge \mathbf{y}|}. \tag{1}$$

Exemple 7. Soient les mots $\mathbf{u} = abab\cdots$ et $\mathbf{v} = aabb\cdots$. On a

$$d(\mathbf{u}, \mathbf{v}) = 1/2.$$

On voit que deux mots sont "d'autant plus proches" qu'ils ont un long préfixe commun.

On vérifiera aisément qu'il s'agit bien d'une *distance*. Autrement dit, il faut vérifier les propriétés suivantes (exercice où il suffit de raisonner sur les préfixes).

Proposition 8. *Pour tous mots infinis* $\mathbf{x}, \mathbf{y}, \mathbf{z} \in A^{\mathbb{N}}$

1. $d(\mathbf{x}, \mathbf{y}) \geq 0$,
2. $d(\mathbf{x}, \mathbf{y}) = 0$ *si et seulement si* $\mathbf{x} = \mathbf{y}$,
3. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$,

2. i.e., on dispose d'une opération binaire interne et partout définie, associative et possédant un neutre. Ainsi, un monoïde dans lequel tout élément est inversible est un groupe.

4. $d(\mathbf{x},\mathbf{z}) \leq d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z})$ *(inégalité triangulaire)*.

Cette distance possède une propriété supplémentaire (exercice du même type que le précédent), elle est *ultramétrique* [3] (on utilise parfois le terme *non-archimédienne*) : elle vérifie

$$\forall \mathbf{x},\mathbf{y},\mathbf{z} \in A^{\mathbb{N}} : \ d(\mathbf{x},\mathbf{z}) \leq \max\{d(\mathbf{x},\mathbf{y}), d(\mathbf{y},\mathbf{z})\}.$$

Ayant à notre disposition un espace métrique $(A^{\mathbb{N}}, d)$, comme dans n'importe quel cours d'analyse, on peut parler de boules, de suites convergentes, etc.

Définition 9. Soit $(\mathbf{x}_n)_{n\geq 0}$ une suite de mots infinis sur A. Cette suite *converge* vers $\mathbf{y} \in A^{\mathbb{N}}$ si

$$\forall \epsilon > 0, \exists N \ : \forall n \geq N, d(\mathbf{x}_n, \mathbf{y}) < \epsilon.$$

Vu (1), cette définition se paraphrase comme suit. Pour toute longueur ℓ, il existe une borne N telle que pour tout $n \geq N$, les mots infinis $\mathbf{x}_n$ ont tous le *même préfixe de longueur* ℓ.

Exemple 10. Dans $\{0,\ldots,9\}^{\mathbb{N}}$, considérons les développements décimaux des premiers convergents du développement en fractions continues [4] de $\pi - 3$:

$\cfrac{1}{7}$	$\underline{14}2857142857142857142857142857142857\cdots$
$\cfrac{1}{7+\cfrac{1}{15}}$	$\underline{1415}0943396226415094339622641509433 96\cdots$
$\cfrac{1}{7+\cfrac{1}{15+\cfrac{1}{1}}}$	$\underline{141592}9203539823008849557522123893 80\cdots$
$\cfrac{1}{7+\cfrac{1}{15+\cfrac{1}{1+\cfrac{1}{292}}}}$	$\underline{14159265}3011902604072261494773729684 00\cdots$

On a d'une part (colonne de gauche), une suite de nombres rationnels convergeant vers le réel $\pi - 3$ et d'autre part, on a une suite de mots infinis convergeant, au sens de la définition 9, vers un mot infini limite. A chaque étape, nous avons souligné le préfixe stabilisé.

On notera que $(A^{\mathbb{N}}, d)$ est un espace métrique complet et compact [18].

Remarque 11. Disposant d'une distance ultramétrique, la topologie associée est intéressante : tout point d'une boule en est le centre, deux boules ont une intersection non vide si et seulement si l'une est incluse dans l'autre, tout triangle est isocèle, etc. Pour s'en convaincre, il suffit d'observer qu'une boule du type

$$\left\{\mathbf{x} \in A^{\mathbb{N}} \mid d(\mathbf{y},\mathbf{x}) \leq 2^{-n}\right\}$$

3. On rencontre notamment ce type de propriété en analyse p-adique. Voir, par exemple, l'excellente introduction [13].

4. Consulter n'importe quel ouvrage classique de théorie des nombres comme [10].

est l'ensemble des mots infinis partageant avec $\mathbf{y}$ le même préfixe de longueur n.

Soit z une lettre n'appartenant pas à A. On peut plonger A^* dans $(A \cup \{z\})^{\mathbb{N}}$ en identifiant le mot fini $w \in A^*$ avec le mot infini $wz^\omega := wzzz\cdots \in (A \cup \{z\})^{\mathbb{N}}$. Cette identification faite, il est licite de parler d'une suite de mots finis convergeant vers un mot infini limite.

Définition 12. Soit $(x_n)_{n\geq 0}$ une suite de mots *finis* sur A. Soit $z \notin A$. Cette suite *converge* vers $\mathbf{y} \in A^{\mathbb{N}}$, si la suite $(x_n z^\omega)_{n\geq 0}$ converge vers $\mathbf{y}$.

Exemple 13. La suite $(x_n)_{n\geq 1}$ de mots finis où x_n est le préfixe de longueur n du développement décimal de $\pi - 3$ converge vers le mot infini correspondant :

$$x_1 = 1,\ x_2 = 14,\ x_3 = 141,\ x_4 = 1415,\ x_5 = 14159,\ x_6 = 141592,\ldots$$

3 Mots morphiques

Pour un alphabet A contenant au moins deux symboles, $A^{\mathbb{N}}$ est un ensemble non dénombrable. En effet, il existe une surjection de $\{0,1\}^{\mathbb{N}}$ dans l'intervalle $[0,1]$. Il suffit de considérer l'application qui à $x_0x_1x_2\cdots$ associe le nombre réel

$$\sum_{i=0}^{+\infty} x_i\ 2^{-i-1}.$$

Cependant, les "algorithmes pouvant engendrer un mot infini de $A^{\mathbb{N}}$" — on entend par là, un programme auquel on fournit un entier n en entrée, par exemple écrit en base 2, et l'algorithme calcule en nombre fini d'opérations le n-ième symbole du mot — forment un ensemble dénombrable. Il n'existe en effet, sur un alphabet fixé, qu'un nombre dénombrable de textes possibles, donc de "programmes". Voir par exemple [21] où Turing s'intéresse aux nombres *calculables*.

Nous allons dès lors nous concentrer sur une classe de mots infinis pour lesquels on dispose d'une représentation succincte (et qui permet de répondre à de nombreuses questions de nature combinatoire). Présentée de façon informelle, la complexité de Chaitin–Kolmogorov classe les mots infinis en fonction de la longueur minimale d'un programme permettant d'en engendrer un préfixe de longueur n. Dans cette section, nous allons définir les mots purement morphiques. Pour ces derniers, en se donnant uniquement les images des lettres de l'alphabet par le morphisme, on peut alors engendrer un préfixe de longueur arbitraire (à partir d'un programme de longueur constante). Par opposition, tout programme engendrant les n premiers symboles d'une suite 'aléatoire' de 0 et de 1 sans la moindre 'structure' doit contenir ces n symboles et a donc une longueur proportionnelle à n.

Soit A^* le monoïde des mots finis sur A. Un *morphisme* est une application $f : A^* \to A^*$ telle que $f(uv) = f(u)f(v)$ pour tous $u, v \in A^*$. En particulier, on doit avoir $f(\varepsilon) = \varepsilon$. Autrement dit, il s'agit d'un homomorphisme compatible avec le produit de concaténation de mots.

Définition 14. Un *morphisme* $f : A^* \to A^*$ *prolongeable sur* a est tel que

— $f(a) = au$ où $u \neq \varepsilon$ est un mot fini et

— $\lim_{n\to+\infty} |f^n(a)| = +\infty$.

Nous profitons de ce premier exemple pour définir la notion de morphisme de longueur constante.

Exemple 15. Considérons un morphisme de *longueur constante*, i.e., les images de chacune des lettres ont même longueur. Ainsi, $t : + \mapsto +-,\ - \mapsto -+$ est prolongeable sur + ou −. Pour tout n, on a directement $|t^n(+)| = 2^n$.

Exemple 16. Soit le morphisme $u : a \mapsto abc,\ b \mapsto ac,\ c \mapsto bac$. Il est prolongeable sur a uniquement. On vérifie que $|u^{n+1}(a)| \geq 2|u^n(a)|$, pour tout n.

Exemple 17. Le morphisme $f : a \mapsto ab,\ b \mapsto a$, appelé morphisme de Fibonacci, est prolongeable sur a. On vérifie, pour tout n, que $|f^n(a)|$ est le n-ième nombre de Fibonacci.

Exemple 18. Les morphismes $g : a \mapsto ba,\ b \mapsto ab$ et $h : a \mapsto ab, b \mapsto \varepsilon$ ne sont pas prolongeables. En effet, on a $|h^n(a)| = 2$ pour tout $n \geq 1$.

Proposition 19. *Soit* $f : A^* \to A^*$ *un morphisme prolongeable sur* a.

1. *Si* $f(a) = au$, *alors pour tout* $n \geq 1$,
$$f^n(a) = au\,f(u)f^2(u)\cdots f^{n-1}(u);$$
2. $f^n(a)$ *est un préfixe de* $f^{n+1}(a)$;
3. *la suite* $(f^n(a))_{n\geq 0}$ *converge vers un mot infini limite noté* $f^\omega(a)$;
4. f *s'étend à* $A^{\mathbb{N}}$ *et le mot infini* $f^\omega(a)$ *est un point fixe de* f.

Démonstration. On procède par récurrence sur n (exercice). Si $\mathbf{x} = x_0x_1x_2\cdots$, on définit $f(\mathbf{x})$ comme la limite de la suite de mots finis $(f(x_0\cdots x_n))_{n\geq 0}$. □

Cette proposition rend licite la définition suivante.

Définition 20. Un mot infini $\mathbf{x}$ est *purement morphique*, s'il existe un morphisme f prolongeable sur une lettre a tel que
$$\mathbf{x} = f^\omega(a) = \lim_{n\to+\infty} f^n(a).$$

Si on reprend l'exemple 15, les premières itérations de t sur + sont

$$\begin{aligned} t(+) &= +- \\ t^2(+) &= +--+ \\ t^3(+) &= +--+-++- \\ t^4(+) &= +--+-++--++-+--+ \\ &\vdots \end{aligned}$$

Le lecteur devrait s'entraîner à calculer, à la main, ces quelques itérations du morphisme t. Le commentaire fait par Cobham, dans son article fondateur [9], n'en serait que renforcé : '*Adding a feedback feature which permits symbols produced at early stages of the generating process to be re-examined at later stages*

increases flexibility and the variety of sequences generable by devices so augmented is substantially richer... Suppose we have generated symbols with index 0 *through* $2k-1$ *and that our left hand points at the k-th of these, our right hand at the last. We observe the symbol at which our left hand is pointing and write with our right the* $2k$*-th and* $(2k+1)$*-st as prescribed. Moving our left hand one symbol to the right, we are in position to repeat the procedure.'*

Autrement dit le symbole + en position 3 donne naissance, par application de t, au facteur $+-$ apparaissant aux positions 6 et 7, etc.

Si on reprend l'exemple 16, les premières itérations de f sur a sont

$$\begin{aligned} u(a) &= abc \\ u^2(a) &= abcacbac \\ u^3(a) &= abcacbacabcbacacabcbac \\ u^4(a) &= abcacbacabcbacacabcbacabcacbacacabc\cdots bac \\ &\vdots \end{aligned}$$

Ici aussi, chaque lettre x donne naissance à un facteur $u(x)$ apparaissant plus loin dans le mot.

4 Retour au problème initial

Avec les notations de la première section, rappelons notre question initiale : *est-il possible de caractériser 'facilement' le signe de* F_n *sur l'intervalle* $I_{n,j}$ *?* En particulier, doit-on déterminer le signe sur chaque sous-intervalle pour déterminer le signe d'un seul d'entre eux ? Doit-on disposer de $sign(I_{n,0}),\dots,sign(I_{n,j-1})$ pour déterminer $sign(I_{n,j})$? Ou encore, doit-on disposer de $sign(I_{n-1,0}),\dots,sign(I_{n-1,2^{n-1}-1})$?

Pour répondre à cette question, l'observation faite à la remarque 2 est primordiale : le signe de F_n sur l'intervalle $I_{n,j}$ détermine entièrement le signe de F_{n+1} sur les intervalles $I_{n+1,2j}$ et $I_{n+1,2j+1}$. Dès lors, pour tout $n \geq 0$, avec la définition de t donnée dans l'exemple 15, la suite des 2^n signes de F_n sur l'intervalle $[0,\pi]$ décomposé en 2^n sous-intervalles est donnée précisément par le mot $t^n(+)$.

Puisque $t^n(+)$ est préfixe de $t^{n+1}(+)$, il nous suffit donc de pouvoir répondre à la question suivante : *si on considère le mot infini* $t^\omega(+) = t_0 t_1 t_2 \cdots$, *peut-on déterminer 'facilement' la valeur de* t_n *?* Par exemple, que vaut $t_{17\,893\,617}$?

Grâce à la remarque 2, nous avons vu qu'il suffit d'itérer le morphisme t de longueur constante 2. On en déduit immédiatement la proposition suivante.

Proposition 21. *Soit* $t^\omega(+) = t_0 t_1 t_2 \cdots = +--+-++-\cdots$ *le mot purement morphique obtenu en itérant le morphisme* $t : + \mapsto +-,\ - \mapsto -+$. *Pour tout* $n \geq 0$, *on a*

$$t_{2n} = t_n, \quad t_{2n+1} = -t_n.$$

Par conséquent, pour déterminer $t_{13} = t_{2.6+1}$, il suffit de connaître $t_6 = t_{2.3+0}$ et d'en changer le signe. Pour cela, il faut donc déterminer $t_3 = t_{2.1+1}$, $t_1 = t_{2.0+1}$ et finalement t_0. Puisque $t_0 = +$, on en tire que $t_1 = -$. De là, $t_3 = +$, puis $t_6 = +$ et enfin, $t_{13} = -$. En y regardant d'un peu plus près, il suffit d'écrire $13 = 2.(2.(2.(2.0+1)+1)+0)+1$ en base 2 : 1101. En lisant de gauche à droite, la suite

de 0 et 1 obtenue, la proposition précédente stipule qu'à chaque lecture d'un chiffre 1, le signe change. Puisqu'on a trois 1 dans l'écriture binaire de 13 et que l'on démarre avec le signe +, le signe de t_{13} est négatif.

Exemple 22. Si on écrit 17 893 617 en base 2, on obtient

$$1000100010000100011110001$$

qui contient 9 symboles 1. Puisque ce mot contient un nombre impair de 1, on conclut de la proposition précédente que l'élément $t_{17\,893\,617}$ est $-$. On est donc en mesure de déterminer le signe de F_n sur n'importe quel sous-intervalle $I_{n,j}$. De plus, l'écriture d'un entier j en base 2 nécessite un nombre de chiffres égal à $\lfloor \log_2 j \rfloor + 1$ et il suffit ici de compter le nombre de 1 apparaissant dans cette écriture.

Nos développements nous amènent à présenter la fonction *somme des chiffres* (ici, en base 2).

Corollaire. *Soit* $t^\omega(+) = t_0 t_1 t_2 \cdots = + - - + - + + - \cdots$ *le mot purement morphique obtenu en itérant le morphisme* $t : + \mapsto +-,\ - \mapsto -+$. *On a*

$$t_n = (-1)^{\sum_{i=0}^{k} c_i}$$

si n *se décompose comme* $\sum_{i=0}^{k} c_i\, 2^i$ *avec* $c_i \in \{0,1\}$ *pour tout* i.

La proposition 21 n'est pas spécifique au morphisme t mais s'étend facilement à tout morphisme de longueur constante [9].

Lemme 23. *Soient* $k \geq 2$ *un entier et* $f : A^* \to A^*$ *prolongeable sur* a *et tel que* $|f(b)| = k$ *pour tout* $b \in A$. *Soit* $\mathbf{x} = f^\omega(a) = x_0 x_1 x_2 \cdots$. *Pour tout* j *tel que* $k^m \leq j < k^{m+1}$, *si on considère la division euclidienne* $j = kq + r$, $k^{m-1} \leq q < k^m$ *et* $0 \leq r < k$, *alors le symbole* x_j *est le* $(r+1)$*-ième symbole apparaissant dans* $f(x_q)$.

De là, on comprend que les écritures en base k permettent de déterminer aisément, i.e., en lisant chaque chiffre de l'écriture une et une seule fois, le j-ième élément d'une suite engendrée par un morphisme de longueur constante k. On pourrait alors introduire la notion de *suite k-automatique* [9]. Un automate fini déterministe (un graphe étiqueté) auquel on fournit l'écriture de j en base k détermine x_j en lisant une et une seule fois chacun des chiffres de l'écriture. L'ouvrage de référence par excellence sur le sujet est [4] mais nous ne voulons pas aller plus loin ici.

5 Une application en combinatoire

La *combinatoire des mots* (classification 68R15 de l'American Mathematical Society) s'intéresse à la structure et aux arrangements apparaissant au sein de mots sur un alphabet fini. Pour une approche historique présentant les problèmes fondateurs de cette discipline, voir [5].

Elle trouve de nombreuses applications. Citons la théorie des nombres et, par exemple, un théorème de transcendance de certains nombres irrationnels dû à Adamczewski et Bugeaud [1], l'algèbre et, par exemple, un théorème à la Skolem–Mahler–Lech en caractéristique p dû à Derken [11], la géométrie discrète (approximation de droites, d'hyperplans, codage de rotations, etc.) et la

dynamique symbolique [6], la théorie ergodique, les systèmes de numération, la vérification, ... Citons, sans être exhaustif, les ouvrages [7, 15].

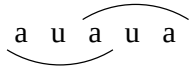
Nous allons ici nous concentrer sur un problème d'évitement déjà étudié par Thue au début du siècle précédent [19, 20] et en relation directe avec le mot $t^\omega(+)$. Commençons par un résultat des plus élémentaires.

Proposition 24. *Sur un alphabet binaire, tout mot de longueur au moins* 4 *contient un carré, i.e., un facteur de la forme* uu, $u \neq \varepsilon$.

Démonstration. Supposons que l'alphabet est $\{a, b\}$ et que le mot débute par a. Essayons de construire un mot sans carré. Ainsi, a doit être suivi par un b. Le mot abb contient un carré. Il faut donc considérer le mot aba. Il n'est pas possible de compléter ce dernier. En effet, $abaa$ contient le carré aa et $abab$ est le carré de ab. □

Cette propriété triviale montre donc que l'apparition d'un carré est *inévitable* sur un alphabet de deux lettres. Qu'en est-il si l'alphabet est de taille 3 ? Sur deux lettres, peut-on éviter l'apparition de cubes ?

Définition 25. Un mot fini de la forme $auaua$ où $u \in A^*$ et $a \in A$ est un *chevauchement* (en anglais, *overlap*).



On remarque que tout chevauchement contient un carré. De même, un cube (i.e., mot de la forme uuu) est un chevauchement particulier.

Nous avons vu que sur un alphabet binaire, tout mot de longueur ≥ 4 contient un carré. Le fait de contenir un chevauchement est une propriété plus forte. Cette propriété est-elle évitable sur deux lettres ? Nous substituerons ici l'alphabet $\{+, -\}$ utilisé jusqu'ici à l'alphabet $\{a, b\}$.

Proposition 26. *Le mot de Thue–Morse, défini comme* $\mathbf{t} = f^\omega(a)$ *où* $f : a \mapsto ab, b \mapsto ba$, *est un mot infini sans chevauchement.*

$$\mathbf{t} = abbabaabbaababbabaababbaabbabaab\cdots$$

La preuve de ce résultat est calquée sur celle présentée dans [16]. Pour des applications (en analyse, en arithmétique et le problème de Prouhet, pour le jeu d'échecs, ...) du mot de Thue–Morse, on lira [3, 17]. Tout comme A^* désigne l'ensemble des mots finis sur l'aphabet A, si X est un ensemble de mots, X^* désigne l'ensemble des mots obtenus en concaténant un nombre fini de mots de X (les répétitions sont autorisées).

Lemme 27. *Soit* $X = \{ab, ba\}$. *Si* x *appartient à* X^*, *alors* axa *et* bxb *n'appartiennent pas à* X^*.

Démonstration. On procède par récurrence sur $|x|$. Si $x = \varepsilon$, il est clair que $aa, bb \notin X^*$. Supposons le résultat vérifié pour les mots de longueur $< n$. Soit $x \in X^*$, un mot de longueur n. Procédons par l'absurde et supposons que $u = axa \in X^*$ (on procède de manière semblable avec bxb). Dans ce cas, $u = abyba$ avec $|y| = |x| - 2$. Puisque $y \in X^*$, on en conclut, par hypothèse de récurrence, que $byb = x$ n'appartient pas à X^*. Ceci est une contradiction. □

En fait, le morphisme de Thue–Morse préserve la propriété d'intérêt.

Lemme 28. *Soient $w \in \{a,b\}^+$ et $f : a \mapsto ab, b \mapsto ba$ le morphisme de Thue–Morse. Si w est sans chevauchement, alors $f(w)$ aussi.*

Démonstration. Montrons que si $f(w)$ possède un chevauchement, alors w aussi. Supposons que $f(w)$ se factorise en

$$f(w) = x\,c\,v\,c\,v\,c\,y, \quad c \in \{a,b\},\ x,v,y \in \{a,b\}^*.$$

Puisque f est 2-uniforme (i.e., $|f(a)| = |f(b)| = 2$), $|f(w)|$ est pair. On remarque que $|cvcvc| = 3 + 2|v|$ est impair. Par conséquent, $|xy|$ est impair.

Montrons à présent que $|v|$ est impair.

— Si $|x|$ est pair, alors x, $cvcv$ et cy appartiennent à $\{ab,ba\}^*$.
Dès lors, si $|v|$ était pair, alors cvc et v appartiendraient à $\{ab,ba\}^*$. Ceci est en contradiction avec le lemme précédent.

— Si $|x|$ est impair, alors xc, $vcvc$ et y appartiennent à $\{ab,ba\}^*$. Si $|v|$ était pair, on aboutirait à la même contradiction.

Nous pouvons à présent conclure, en discutant une fois encore sur la parité de $|x|$.

— Si $|x|$ est pair, alors, puisque $|v|$ est impair, on a

$$f(w) = \underbrace{x}_{\text{pair}}\ \underbrace{c\ \overbrace{v}^{\text{impair}}}_{\text{pair}}\ \underbrace{cv}_{\text{pair}}\ cy \in \{ab,ba\}^*$$

et x, cv, cy appartiennent à $\{ab,ba\}^*$. Il existe r,s,t tels que $f(r) = x$, $f(s) = cv$, $f(t) = cy$ et

$$w = rsst.$$

Or, $f(s)$ et $f(t)$ débutent par la même lettre, donc s et t aussi (vu la définition de f). Par conséquent, sst débute par un chevauchement.

— Si $|x|$ est impair, alors, puisque $|v|$ est impair, on a

$$f(w) = \underbrace{xc\ \overbrace{v}^{\text{impair}}}_{\text{pair}}\ \underbrace{c}_{\text{pair}}\ \underbrace{vc}_{\text{pair}}\ y \in \{ab,ba\}^*$$

et il existe r,s,t tels que $f(r) = xc$, $f(s) = vc$, $f(t) = y$. La conclusion est identique.

□

Nous pouvons à présent démontrer la proposition 26.

Démonstration. Supposons que **t** possède un chevauchement. En particulier, ce chevauchement apparaît dans le préfixe $f^k(a)$ pour un certain k. Or, a étant sans chevauchement, le lemme précédent stipule que $f(a)$ est sans chevauchement et donc, en itérant, $f^k(a)$ ne peut posséder de chevauchement. □

La proposition 26 va nous permettre de construire un mot infini sur trois lettres évitant les carrés.

Remarque 29. Soit **r** un mot infini sur $\{a,b\}$ sans chevauchement et commençant par a. Alors, **r** se factorise de manière unique sous la forme [5] $\mathbf{r} = y_1 y_2 \cdots$ où pour tout $i \geq 1$, $y_i \in \{a, ab, abb\}$. En effet, **r** ne contenant aucun cube, il ne peut contenir le facteur *aaa* ou *bbb*.

Soit le morphisme $g : \{a,b,c\}^* \to \{a,b\}^*$ défini par

$$g : \begin{cases} a \mapsto abb \\ b \mapsto ab \\ c \mapsto a \end{cases} .$$

Si **r** un mot infini sur $\{a,b\}$ sans chevauchement et débutant par a, alors il existe un unique mot infini **s** sur $\{a,b,c\}$ tel que $g(\mathbf{s}) = \mathbf{r}$.

Proposition 30. *Soit* **t** *un mot infini sur* $\{a,b\}$ *sans chevauchement et débutant par a (comme le mot de Thue–Morse). Soit* **s** *l'unique mot infini* $\{a,b,c\}$ *tel que* $g(\mathbf{s}) = \mathbf{t}$. *Alors,* **s** *est un mot infini sur trois lettres sans carré.*

Démonstration. Supposons que **s** contienne un carré : $\mathbf{s} = x\,u\,u\,\sigma\,y$ avec u non vide, σ une lettre et y un mot infini. Alors, $g(\mathbf{s})$ contient le facteur $g(u)g(u)g(\sigma)$ qui débute par un chevauchement car $g(u)$ et $g(\sigma)$ débutent par la même lettre. □

Voici la factorisation fournie par le résultat précédent :

$$\mathbf{t} = abb|ab|a|abb|a|ab|abb|ab|a|ab|abb|a|abb|ab|a|a\cdots$$

$$\mathbf{s} = abcacbabcbacabc\cdots$$

Remarque 31. On peut montrer que le mot **s** donné ci-dessus est purement morphique ; il s'obtient comme $\mathbf{s} = \varphi^\omega(a)$ où $\varphi(a) = abc$, $\varphi(b) = ac$ et $\varphi(c) = b$, est sans carré. Ce morphisme φ est parfois appelé *morphisme de Hall*.

6 Répétitions abéliennes

Pour terminer cette note, mentionnons une généralisation possible des résultats d'évitement de la section précédente.

Définition 32. Un *carré abélien* est un mot uv où v est un anagramme de u, i.e., v est obtenu en permutant les lettres de u. Par exemple, *abcbca* est un carré abélien.

De façon analogue à la proposition 24, on peut montrer (exercice de programmation) que sur un alphabet de trois lettres, tout mot suffisamment long contient un carré abélien. Ainsi, les mots les plus longs sans carré abélien sont 0102010 et 0102101.

Dans une liste de 28 problèmes posés par Erdős [12], le dernier d'entre eux est celui-ci : *"Let* $N(k)$ *be the least number* N *with the property that each sequence* $\{s_n\}_{n=1}^N$ *of numbers taken from the set* $\{1,\ldots,k\}$ *contains two adjacent blocks such that*

5. On remarquera que $Y = \{a, ab, abb\}$ est un code, i.e., tout mot de Y^* possède une unique factorisation comme concaténation d'éléments de Y.

each is a rearrangement of the other. My earliest conjecture, that $N(k) = 2^k - 1$, *has been disproved by Bruijn and myself. It is not even known whether* $N(4) < \infty$."

Ce problème est resté ouvert pendant plus de trente ans. Keränen lui a apporté une réponse positive à l'aide d'un mot purement morphique [14].

Théorème 33. *Le mot infini* $k^\omega(a)$ *obtenu en itérant le morphisme prolongeable défini par*

$$a \mapsto abcacdcbcdcadcdbdabacabadbabcbdbcbacbcdcacbabd$$
$$abacadcbcdcacdbcbacbcdcacdcbdcdadbdcbca;$$

$$b \mapsto bcdbdadcdadbadacabcbdbcbacbcdcacdcbdcdadbdcbca$$
$$bcbdbadcdadbdacdcbdcdadbdadcadabacadcdb;$$

$$c \mapsto cdacabadabacbabdbcdcacdcbdcdadbdadcadabacadcdb$$
$$cdcacbadabacabdadcadabacabadbabcbdbadac;$$

$$d \mapsto dabdbcbabcbdcbcacdadbdadcadabacabadbabcbdbadac$$
$$dadbdcbabcbdbcabadbabcbdbcbacbcdcacbabd;$$

ne contient aucun carré abélien.

Dans la même veine, nous terminerons en citant le résultat suivant [8]. Un *cube additif* est un mot sur un alphabet d'entiers naturels de la forme uvw tel que $|u| = |v| = |w|$ et les sommes respectives des symboles constituant u, v et w sont identiques. Par exemple, 041322 est un cube additif, $0 + 4 = 1 + 3 = 2 + 2$.

Théorème 34. *Soit le morphisme* $\varphi : 0 \mapsto 03,\ 1 \mapsto 43,\ 3 \mapsto 1,\ 4 \mapsto 01$. *Le mot infini*

$$\varphi^\omega(0) = 0314301103434303101101103143034343034\cdots$$

ne contient aucun cube additif.

Ces deux théorèmes montrent, une fois encore, que des mots purement morphiques permettent de répondre à des questions combinatoires difficiles.

7 Bibliographie

[1] B. Adamczewski, Y. Bugeaud On the complexity of algebraic numbers I. Expansions in integer bases, *Ann. Math*, **165**, 547–565, (2007).

[2] J.-P.. Allouche, M. Mendès France, Euler, Pisot, Prouhet-Thue-Morse, Wallis and the duplication of sines, *Monatsh. Math.*, **155**, 301–315, (2008).

[3] J.-P.. Allouche, J. Shallit, The ubiquitous Prouhet-Thue-Morse sequence, Sequences and their applications (Singapore, 1998), 1—16, *Springer Ser. Discrete Math. Theor. Comput. Sci.*, Springer, London, (1999).

[4] J.-P.. Allouche, J. Shallit, *Automatic Sequences : Theory, Applications, Generalizations*, Cambridge University Press, (2003).

[5] J. Berstel, D. Perrin, The origins of combinatorics on words, *European J. Combin.*, **28**, 996–1022, (2007).

[6] V. Berthé, Discrete Geometry and Symbolic Dynamics, The Kiselmanfest : An International Symposium in Complex Analysis and Digital Geometry, (2006), Uppsala, Sweden.

[7] V. Berthé, M. Rigo, (Eds.) *Combinatorics, Automata and Number Theory*, Encycl. of Math. and its Appl. **135**, Cambridge University Press, (2010).

[8] J. Cassaigne, J. D. Currie, L. Schaeffer, J. Shallit, Avoiding three consecutive blocks of the same size and same sum, arXiv:1106.5204

[9] A. Cobham, Uniform tag sequences, *Math. Systems Theory*, **6**, 164–192, (1972).

[10] H. Davenport, *Higher arithmetic. An introduction to the theory of numbers*. Dover Publications, Inc., New York, (1983).

[11] H. Derksen, A Skolem-Mahler-Lech theorem in positive characteristic and finite automata, *Invent. Math.*, **168**, 175–224, (2007).

[12] P. Erdős, Some unsolved problems, *Michigan Math. J.*, **4**, 291–300, (1957).

[13] F. Gouvêa, *p-adic Numbers - An Introduction*, Universitext, Springer, (1997).

[14] V. Keränen, Abelian squares are avoidable on 4 letters, *Lecture Notes in Comput. Sci.*, **623**, 41–52, (1992).

[15] D. Lind, B. Marcus *An Introduction to Symbolic Dynamics and Coding*, Cambridge Univ. Press, (1996).

[16] M. Lothaire, *Combinatorics on Words*, Cambridge University Press, (1983).

[17] M. Rigo, Le problème de Prouhet, *Losanges*, **19**, 42–53, (2012). Disponible en ligne à http://hdl.handle.net/2268/129326

[18] M. Rigo, *Formal Languages, Automata and Numeration Systems : Introduction to Combinatorics on Words*, vol. 1, ISTE-Wiley, (2014).

[19] A. Thue, Über unendliche Zeichenreihen, *Norske vid. Selsk. Skr. Mat. Nat. Kl.*, **7**, 1–22, (1906).

[20] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Norske vid. Selsk. Skr. Mat. Nat. Kl.*, **1**, 1–67, (1912).

[21] A. Turing, On computable numbers, with an application to the Entscheidungsproblem, *Proc. London Math. Soc.*, **42**, 230–265, (1937).