J. BERSTEL

Mots infinis

Publications du Département de Mathématiques de Lyon, 1984, fascicule 6B
« Théorie des langages et complexité des algorithmes », , p. 89-102

<http://www.numdam.org/item?id=PDML_1984___6B_A4_0>

MOTS INFINIS

par J. BERSTEL

I. Introduction

Depuis que l'informatique existe, elle ne traite pas que des nombres, mais aussi des mots. Ceci se reflète dans l'étude théorique qui accompagne l'évolution de l'emploi des ordinateurs. On constate toutefois, en feuilletant la littérature, que la plupart des investigations sur les mots sont consacrées aux suites finies de symboles, et que l'étude des mots infinis n'a eu, dans le temps, qu'un impact assez faible, comparé à la théorie plus classique des mots finis. Depuis plusieurs années, on assiste maintenant à une recrudescence de travaux sur les mots infinis, et il y a diverses raisons pour cela.

L'une des notions les plus difficiles à formaliser de manière satisfaisante est la notion de calcul. Et il existe un nombre considérable de développements théoriques qui lui sont consacrés. Ne citons que deux situations où ce concept est employé : les dérivations dans les grammaires et les calculs dans les programmes ou schémas de programmes. Dans les deux cas, on s'aperçoit que l'introduction d'objets infinis (dérivations infinies respectivement calculs infinis) permet de rendre compte de manière nettement plus satisfaisante des phénomènes rencontrés. Formellement, ces objets sont des mots infinis.

Mais en fait, le concept de mot infini est courant en mathématique au moins depuis le début du siècle, ou certaines études logico-combinatoires en ont fait leur objet de prédilection.

Le but de ces pages est de présenter quelques propriétés combinatoires des mots infinis. Ces propriétés n'ont certes pas d'influence directe sur l'évolution immédiate de l'informatique, mais elles font partie de ces soubassements théoriques qui mettent en perspective l'activité plus militante des acteurs en première ligne.

Considérons, pour commencer, le mot infini que voici, bien connu sous le nom de mot de Fibonacci :

$$f = abaababaabaababaababa...$$

Le type de questions que nous voulons poser ici sont :

- Comment construire un mot infini ? Il s'agit de donner un procédé fini explicite et le plus simple possible pour obtenir un mot infini. Les morphismes itérés et les tag-systèmes de Cobham sont des mécanismes bien adaptés. Le mot de Fibonacci s'obtient ainsi.

- Quels sont les facteurs d'un mot infini ? Les blocs qui apparaissent dans un mot infini construit d'une manière particulière ne sont pas quelconques. On se demande alors si on peut les compter, et si on peut les décrire. Dans le mot de Fibonacci par exemple, il y a exactement $n+1$ facteurs de longueur n pour tout entier n.

- Enfin, quelles sont les régularités d'un mot infini ? Là, on s'intéresse aux répétitions qui figurent dans le mot, et on parlera de mots sans carré. Le mot de Fibonacci contient des carrés, et même des cubes, mais pas de puissance quatrième.

Les résultats présentés ici ne sont pas nouveaux. Certains datent du début du siècle, où A. Thue a consacré plusieurs articles longs et substantiels aux mots infinis. Depuis quelques années, l'intérêt pour ces propriétés combinatoires croît, et de nombreux problèmes restent toujours ouverts.

II. Construire

Mot de Fibonacci

Le mot de Fibonacci dont nous avons donné le début ci-dessus peut être construit de plusieurs manières. En voici deux :

On considère la suite de mots (finis) définie par récurrence comme suit sur un alphabet à deux lettres a et b :

$$f_0 = a \qquad f_1 = ab \qquad f_{n+2} = f_{n+1} f_n$$

Les premiers termes de la suite ainsi obtenue sont :

$$f_0 = a$$

$$f_1 = ab$$

$$f_2 = aba$$

$$f_3 = abaab$$

$$f_4 = abaababa$$

$$f_5 = abaababaabaab$$

...

Le mot de Fibonacci est alors, par définition, le mot obtenu en "passant à la limite" dans cette suite : c'est le mot infini (unique) dont les mots de la suite ci-dessus sont des facteurs gauches. Par commodité, on écrit

$$f = \lim_{n \to \infty} f_n$$

Une deuxième manière d'obtenir le mot de Fibonacci est plus systèmatique, même si elle n'est qu'une reformulation du premier procédé. On considère, pour cela le morphisme

$$\phi : \{a, b\}^* \to \{a, b\}^*$$

défini par

$$\phi(a) = ab \; ; \; \phi(b) = a.$$

On calcule alors les itérés successifs de ce morphisme sur la lettre a.

$$\phi^0(a) = a$$
$$\phi^1(a) = ab$$
$$\phi^2(a) = aba$$
$$\phi^3(a) = abaab$$
$$\phi^4(a) = abaababa$$

...

On constate facilement que les mots ainsi obtenus sont les mêmes que précédemment, plus précisément que

$$\phi^n(a) = f_n$$

Ainsi

$$f = \lim_{n \to \infty} \phi^n(a),$$

ce que l'on note aussi en écrivant

$$f = \phi^{\omega}(a)$$

On dit que le mot est obtenu au moyen du morphisme itéré ϕ .

Mot de Thue-Morse

Le deuxième exemple de mot infini, bien connu lui aussi, est appelé mot de Thue ou mot de Morse. Il a été trouvé en effet indépendamment par Thue en 1906 [22] et par Morse en 1921 [19].
C'est le mot :

$$m = 01101001100101101001...$$

obtenu (par exemple) au moyen du morphisme μ suivant :

$$\mu(0) = 01 \quad ; \quad \mu(1) = 10.$$

En itérant à partir de 0, on construit successivement :

$$\mu(0) = 01$$
$$\mu^2(0) = 0110$$
$$\mu^3(0) = 01101001$$
$$\ldots$$

d'où

$$m = \mu^{\omega}(0)$$

Mots des carrés

Ce troisième exemple est défini de façon différente. Ce mot

$$q = 1\ 1\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ ..$$
$$0\ 1 \quad 4 \quad 9 \quad 16$$

possède un "1" en position n si n est un carré, et un "0" sinon.
On peut, à propos de ce mot, montrer facilement :

OBSERVATION : Il n'existe pas de morphisme itéré qui engendre le mot infini q.
Par conséquent, ce mot q montre une limite au mécanisme de construction par morphisme itéré. Et en effet, ce procédé est fort contraint, et il est notamment très peu souple vis-à-vis de transformations pourtant benignes, comme la suppression ou l'adjonction d'une lettre en début d'un mot infini. Il existe un autre modèle de construction de mots infinis, présenté sous une forme très

voisine par Cobham [7] et inspiré en fait de travaux de Post dans les années 20 (voir à ce sujet le livre de Minsky [18]). Un tag-système est composé de deux morphismes, disons ϕ et α , le premier opérant sur un alphabet auxiliaire selon le mode des morphismes itérés, le deuxième n'étant employé qu'à la fin, pour "effacer" les traces des lettres temporaires. Pour le mot des carrés, les deux étapes se présentent comme suit :

Soit

$$\phi : \{0, 1, 2\}^* \longrightarrow \{0,1,2\}^*$$

défini par

$$\phi(0) = 0$$
$$\phi(1) = 001$$
$$\phi(2) = 21$$

Les itérations successives donnent

$$\phi(2) = 21$$
$$\phi^2(2) = 21001$$
$$\phi^3(2) = 2100100001$$

On obtient donc un mot infini, soit Q, qui a la forme

$$Q = \lim_{n \to \infty} \phi^n(2) = 2100100001000000001\ldots$$

Soit maintenant

$$\alpha : \{0,1,2\}^* \longrightarrow \{0,1\}^*$$

$$\alpha(0) = 0 \; ; \; \alpha(1) = \alpha(2) = 1.$$

Alors on a

$$q = \alpha(Q) = \alpha(\phi^{\omega}(2)).$$

Le dernier des trois exemples a été spécifié de façon différente : on est parti d'une description des lettres que l'on s'attendait à trouver aux diverses positions dans le mot, et on a donné un moyen de construire ce mot. Réciproquement, si l'on est en présence d'un mot défini par un morphisme itéré ou par un tag-système, on peut se demander quelles sont les positions où apparaît une lettre donnée. De manière différente et un peu simplifiée, on peut formuler la question comme suit :

Etant donné un mot infini x sur l'alphabet $\{0,1\}$, quel est son support, i.e. quels sont les entiers n tels que la n-ième lettre x(n) du mot x vaut 1 ?

Pour le mot des carrés, c'est l'ensemble des carrés.

Pour le mot de Thue-Morse, la solution appartient au folklore : on a $m(n)=1$ ssi le nombre de "1" figurant dans l'écriture binaire de n est impair. C'est là d'ailleurs un exemple très particulier d'un résultat dû à Cobham [7] qui dit que dans le cas d'un morphisme itéré ou d'un tag-système uniforme de module k (les images des lettres du morphisme qui s'itère ont toutes longueur k), les supports sont bien connus : ce sont exactement les ensembles de nombres k-reconnaissables. (Voir aussi Christol et al. [6]).

Pour le mot de Fibonacci enfin, on peut monter que la n-ième lettre vaut "a" ssi

$$n+1 \in A = \{\lfloor \phi k \rfloor - 1 : k > 0\}$$

où bien sûr

$$\phi = \frac{1+\sqrt{5}}{2}$$

Ces cas particuliers sont, il faut bien le dire, assez exceptionnels : en général, on ne connait aucun moyen de cette nature pour décrire les positions des lettres dans un mot, ni réciproquement de procédé itératif simple pour engendrer un mot avec support prescrit. Il convient de citer, en plus des ensembles reconnaissables de nombres, deux extensions des propriétés :

Soit P un polynôme avec $P(\mathbb{N}) \subset \mathbb{N}$. Soit p le mot infini sur $\{0,1\}$ tel que $p(n) = 1$ ssi $n \in P(\mathbb{N})$. Alors p est engendré par un tag-système.

Deuxièmement, notons $bin(n)$ l'écriture binaire de n, et notons comme d'usage $|bin(n)|_w$ le nombre d'occurences du mot w comme facteur (bloc) dans $bin(n)$. Par exemple,

$$bin(53) = 110101$$

et si $w = 101$, on a

$$|bin(53)|_w = 2$$

Soit m_w le mot infini sur $\{0,1\}$ tel que

$$m_w(n) = 1 \quad \text{ssi} \quad |bin(n)|_w \text{ est impair}$$

Alors m est engendré par un tag-système (Christol et al. [6]).
Observons que m_1 est la suite de Thue-Morse, et

$$m_{11} = 00010010000111\ldots$$

est connu sous le nom de suite de Rudin-Shapiro (ibid.).

III. Facteurs

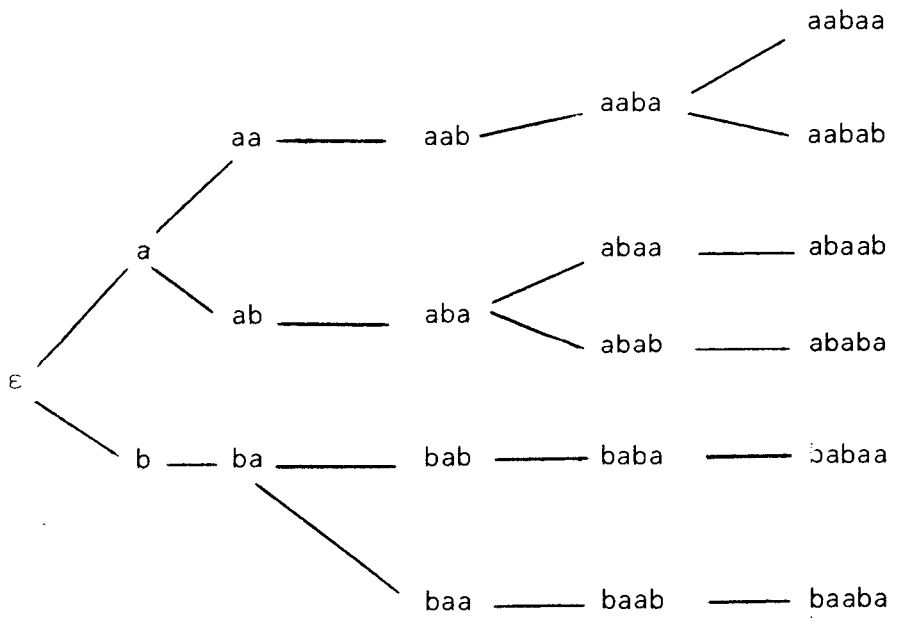
Un facteur d'un mot infini est un mot fini qui apparaît comme bloc (facteur) à l'intérieur de ce mot. Si aucune condition n'est imposée au mot, n'importe quel ensemble de facteurs est possible (sous la condition bien sûr qu'il contiennent les facteurs de ses éléments).
En particulier, tout mot peut apparaître comme facteur dans ce mot.
Il en est ainsi du mot

12345678901011121314151617181920...

forme de la juxtaposition des entiers naturels en notation décimale. En revanche, un mot particulier, et il en est ainsi notamment des mots obtenus en itérant un morphisme ou par tag-système, ne possède que des facteurs d'une forme particulière. Considérons comme exemple le mot de Fibonacci

f = abaababaabaababaabab...

Visiblement, le mot bb n'est pas facteur, ni d'ailleurs le mot aaa.
Plus précisément, on a le tableau suivant des facteurs du mot de Fibonacci, ou chaque mot est relié au mot obtenu en le privant de sa dernière lettre



Introduisons la

NOTATION : Soit x un mot infini. On note

$$P(x,n)$$

le nombre de facteurs de x de longueur n.

OBSERVATION : On a $P(f,n) = n+1$ pour tout $n \geq 0$.

Cette observation appelle plusieurs remarques. Tout d'abord, le mot de Fibonacci a peu de mots. En effet, le nombre $P(x,n)$ est majoré, dans le cas d'un mot écrit sur deux lettres, par 2^n. En fait, le mot étant engendré par un morphisme, il tombe sous le coup du résultat suivant, dû a Ehrenfeucht, Lee et Rosenberg [11] :

PROPOSITION : Soit x un mot infini engendré par un tag-système.
Alors on a $P(x,n) = O(n^2)$.

D'autre part, on a des renseignements assez précis sur le nombre minimum de facteurs que doit contenir un mot infini non trivial, et ceci quel que soit la façon de l'engendrer. Le résultat suivant est extrait, dans sa formulation, d'un article de Coven, Hedlund [5] :

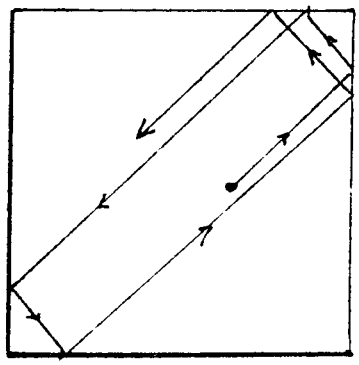
PROPOSITION : Soit x un mot infini. Les conditions suivantes sont équivalentes :

(i) $P(x,n) \leq n$ pour un $n \geq 1$;
(ii) $\{ P(x,n) : n \geq 1 \}$ est borné ;
(iii) x est ultimement périodique.

COROLLAIRE : Si x n'est pas ultiment périodique, alors $P(x,n) \geq n+1$ pour tout n.

En d'autres termes, le mot de Fibonacci est l'un des mots pour lesquels la fonction P est minimale. On peut alors se demander quels sont les mots infinis x, disons sur l'alphabet $\{0,1\}$ qui sont "minimaux" dans ce sens, c'est-à-dire qui vérifient $P(x,n)=n+1$ pour tout n. Cette question a été bien étudiée (Coven, Hedlund [9] , Morse, Hedlund[13]). Considérons le jeux de billard suivant : Dans un carré de côté 1, on choisit un point et on lance une bille selon un angle déterminé. Elle vient se heurter aux côtés, et est réfléchie sans frottement ni autre effet. Pour chaque choc sur une

parois horizontale, on écrit un "1" et pour chaque choc vertical un "0". On obtient ainsi un mot infini sur l'alphabet $\{0,1\}$ qui, si l'angle de lancement est irrationnel modulo π, est un mot minimal.



IV. Régularités

Nous nous intéressons ici aux régularités concernant des facteurs consécutifs dans un mot infini. Plus précisément, on appelle carré un mot de la forme uu (et de même cube, puissance k-ième...) et chevauchement un mot de la forme

$$vuvuv, \text{ avec } v \neq \varepsilon .$$

Comme exemple, considérons à nouveau nos deux suites. Le mot de Fibonacci contient des cubes, mais pour en trouver, il faut déjà le développer assez loin :

$$f = abaababaabaababaababaabaababa\ldots$$

OBSERVATION (Karhumäki [15]). - Le mot de Fibonacci f ne contient pas de puissance quatrième.

Pour le mot de Thue-Morse m, la situation est différente. Bien sûr, ce mot contient des carrés comme tout mot assez long sur 2 lettres.

THÉORÈME (Thue [22]). - Le mot de Thue-Morse m est sans facteur chevauchant.

A partir du mot de Thue-Morse m, A. Thue a construit un mot sans carré sur un alphabet à 3 lettres de la manière suivante :

comme m est sans chevauchement, a fortiori sans cube, chaque lettre "0" de m est suivie d'au plus deux lettres "1" consécutives. On établit alors le codage :

$$a \longrightarrow 011$$
$$b \longrightarrow 01$$
$$c \longrightarrow 0$$

et on note t le mot m réécrit selon ce décodage

$$t = abcacbabcbac \ldots$$

THÉORÈME (Thue [23]). - Le mot t est sans carré.

Une intéressante question, maintenant que nous savons qu'il existe une infinité de mots sans carré, est la suivante : existe-t-il "beaucoup" de mots sans carré ? Plus précisément, fixons un alphabet à trois lettres et notons c(n) le nombre de mots sans carré sur cet alphabet. On a le résultat surprenant que voici :

THÉORÈME (Brandenburg [2]). - On a, pour tout $n \geqslant 2$,

$$6 \cdot (1.032)^n \leqslant c(n) \leqslant 6 \cdot (1.38)^n$$

Ainsi donc y a-t-il croissance exponentielle du nombre de mots sans carré. (Un autre article de Brinkhuis [3] est un peu moins précis). Ce résultat prend tout son relief comparé à l'estimation analogue pour les mots sans facteur chevauchant :

THÉORÈME (Restivo, Salemi [20]). - Le nombre $\gamma(n)$ de mots sans facteur chevauchant de longueur n sur deux lettres vérifie

$$\gamma(n) \leqslant C\, n^{\log_2 15}$$

La construction du mot t a partir du mot m dont nous avons parlé prend un intérêt nouveau à la lumière de cet énoncé : la construction associe en fait a tout mot sans chevauchement sur $\{0,1\}$ commençant par 0 un mot sans carré sur $\{a,b,c\}$. Les deux théorèmes montrent que cette correspondance ne couvre

qu'une faible partie des mots sans carré.

V. Morphismes sans carré

Un morphisme $h : A^* \dashrightarrow B^*$ est un morphisme sans carré si pour tout mot sans carré w dans A^*, l'image $h(w)$ de w par le morphisme h est elle aussi sans carré. Un tel morphisme, si c'est un endomorphisme, est utile pour la construction de mots sans carré par itération : soit en effet $h : A^* \longrightarrow A^*$ un morphisme tel que $h(a)$ commence par la lettre a. Si h est un morphisme sans carré, tous les mots $h^n(a)$, pour $n \geqslant 1$, sont sans carré, et de même $h^{\omega}(a)$.

Exemple : Soit h le morphisme défini par

$$h(a) = abc$$
$$h(b) = ac$$
$$h(c) = b$$

(Voir par exemple Istrail [14]). Comme le suggèrent les premiers mots :

$$h(a) = abc$$
$$h^2(a) = abcacb$$
$$h^3(a) = abcacbabcbac$$

On a

$$t = \lim_{n \to \infty} h^n(a) .$$

Pourtant, h n'est pas un morphisme sans carré, puisque

$$h(aba) = abcacabc$$

contient le carré $(ca)^2$.

De toute façon, un morphisme sans carré ne peut pas être si simple. C'est ce qu'a prouvé A. Carpi [4] : Il a montré qu'un endomorphisme h sans carré sur un alphabet à trois lettres a,b,c doit vérifier l'inégalité $|h(a)| + |h(b)| + |h(b)| \geqslant 18$. Ainsi l'exemple de Thue [23]

$$g(a) = abcab$$
$$g(b) = acabcb$$
$$g(c) = acbcacb$$

était déjà optimal...

Du coup, l'existence de morphismes sans carré n'est plus tellement évidente. Le vrai problème est, comme un peu de réflexion le montre, la construction d'un morphisme sans carré d'un alphabet à 4 lettres sur un alphabet à 3 lettres. Une fois ceci fait, on peut construire par composition (les morphismes sans carré sont stables par composition) des morphismes sans carré de tout alphabet dans tout alphabet (de taille $\geqslant 3$). Un premier morphisme sans carré d'un alphabet à 4 lettres sur un alphabet à 3 lettres a été donné par Bean, Ehrenfeucht et Mc Nulty [1] . Des morphismes plus courts sont dans la thèse de Crochemore [10] . Voir aussi Brandenburg (cite).

Nous terminons par le problème de décider si un morphisme est sans carré, cube etc. On a le

THÉORÈME (Crochemore [8,9]). - Soit $h : A^* \longrightarrow B^*$ un morphisme. Alors h est sans carré ssi

(1) Pour tout mot x de longueur 3 sans carré, l'image $h(x)$ est sans carré.

(2) Aucun des mots $h(a)$ $(a \in A)$ ne contient de précarré interne.

Un mot u, facteur de $h(a)$ est appelé un précarré interne si l'on peut prolonger a en ax de telle sorte que $h(ax)$ contient le carré uu. (Et bien sûr, la même condition à gauche). Ce résultat est une analyse très fine des possibilités de chevauchement à l'intérieur des mots. Il implique en particulier qu'il est décidable si un morphisme est sans carré.

COROLLAIRE (Crochemore [8,9]). - Soit $h : A^* \longrightarrow B^*$ un morphisme, avec A un alphabet à 3 lettres. Alors h est sans carré ssi $h(x)$ est sans carré pour tous les mots x sans carré de longueur 5.

Pour les morphismes sans cube, les choses ne sont pas encore aussi avancées. On ne possède que des résultats partiels. Citons

THEOREME (Karhumäki [15]). - Il est décidable si un morphisme

$$h : \{a,b\}^* \to B^*$$

est sans cube.

[1] D. Bean, A. Ehrenfeucht, G. Mc Nulty, Avoidable patterns in strings of symbols,
Pacific J. Math. 85(1979), 261-294

[2] F. Brandenburg, Uniformly growing k-th power-free homomorphisms,
Theor. Comput. Sci. 23(1983), 69-82

[3] J. Brinkhuis, Non-repetitive sequences on three symbols,
Quart. J. Math. Oxford(2) 34(1983), 145-149.

[4] A. Carpi, On the size of a square-free morphism on a three letter alphabet, prepublication, 1983.

[5] E. Coven, G. Hedlund, Sequences with minimal block growth,
Math. Syst. Theory 7(1973), 138-153.

[6] C. Christol, T. Kamae, M. Mendès-France, G. Rauzy, Suites algébriques, automates et substitutions,
Bull. Soc. Math. France 108(1980), 401-419.

[7] A. Cobham, Uniform tag sequences,
Math. Systems. Th. 6(1972), 164-192.

[8] M. Crochemore, Sharp characterizations of squarefree morphisms,
Theor. Comput. Sci. 18(1982), 221-226.

[9] M. Crochemore, Mots et morphismes sans carré,
Annals of Discr. Math. 17(1983), 235-245

[10] M. Crochemore, Régularités évitables, Thèse d'Etat, Rouen 1983.

[11] A. Ehrenfeucht, K. Lee, G. Rozenberg, Subword complexities of various classes of deterministic developmental languages without interaction,
Theor. Comput. Sci. 1(1975), 59-75.

[12] A. Ehrenfeucht, G. Rozenberg, On the subword complexity of square-free DOL languages
Theor. Comp. Sci. 16(1981), 25-32.

[13] G. Hedlund, M. Morse, Symbolic Dynamics II. Sturmian trajectories,
American J. Math 62(1940), 1-42.

[14] S. Istrail, On irreductible languages and non rational numbers,
Bull. Mat. Soc. Sci. Mat. R. S. Roumanie 21(1977), 301-308.

[15] J. Karhumäki, On cube-free w-words generated by binary morphisms,
Discr. Appl. Math 5(1983), 279-297.

[16] M. Lothaire, "Combinatorics on Words",
Addison-Wesley 1983.

[17] M. Main, Permutations are not context-free : an application of the
'Interchange Lemma'
Inf. Proc. Letters 15(1982), 68-71.

[18] M. Minsky, Computations: finite and infinite machines, Prentice-Hall 1967.

[19] M. Morse, Recurrent geodesics on a surface of negative curvature,
Trans. Amer. Math. Soc. 22(1921), 84-100

[20] A. Restivo, S. Salemi, On weakly square free words,
Inf. Proc. Letters, to appear.

[21] R. Shelton, R. Soni, Aperiodic words on three symbols, I, II, III,
J. reine angew. Math. 321(1981), 195-209, 327(1981), 1-11,
330(1982), 44-52.

[22] A. Thue, Über unendliche Zeichenreihen,
Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania (1906),
1-22.

[23] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser
Zeichenreihen,
Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania (1912), 1-67.