Overlap-Free Words

Julien Cassaigne
ENS Paris

November 29, 1993

[summary by Philippe Dumas]

An overlap-free word is a word without overlapping of two distinct occurrences of a factor. As an example the word *abaab* is overlap-free, but the word *abbbabbabaab* is not because the factor *bbab* occurs twice and the occurrences overlap by *b*. The first result about these words is due to Axel Thue [10]. Indeed the infinite word now referred to as the Thue-Morse word has no overlapping factor [8, p. 23]. Recall that this word may be defined as a fixed point $t = \vartheta^\omega(a)$ of the substitution ϑ which maps a and b onto ab and ba respectively. The proof depends upon the fact that $\vartheta(w)$ has no overlapping factor if the word w does not.

We want to study the number u_n of overlap-free words of length n over a two letter alphabet $\{a, b\}$. All the factors of the Thue-Morse word t are overlap-free words. This yields the inequality $u_n \geq t_n$, where t_n is the number of factors of t with length n. This number t_n is known to be [2, 5]

$$t_n = \begin{cases} 4n - 2.2^k & \text{if} \quad 2.2^k \leq n \leq 3.2^k, \\ 2n + 4.2^k & \text{if} \quad 3.2^k \leq n \leq 4.2^k. \end{cases}$$

On the other hand, Restivo and Salami [9] have shown the asymptotic inequality[1]

$$u_n \preccurlyeq n^{\log_2 15}.$$

Kobayashi [7] improved these results by the estimate

$$n^{1.155} \preccurlyeq u_n \preccurlyeq n^{1.587}.$$

Overlap-free words which can be extended to infinity in both directions are factors of the Thue-Morse word. Counting the words extensible to the right gives the lower bound. Thanks to ϑ, one can build overlap-free words of length $2n$ from overlap-free words of length n, hence the upper bound. Carpi [3] gave a way to obtain upper bounds n^r arbitrarily near the optimal value, but gave no numerical result. In addition he pointed out that one can compute the sequence u_n by a finite automaton.

1. Linear representation

Decomposition. All the cited results rely upon a decomposition of the overlap-free words. Apart from a set S of seventy-six little words, every word which is overlap-free and of length greater than 3, or which has a unique overlapping by only one letter (J. Cassaigne uses the term of minimal overlap for these words $xuxux$, where x is a letter and u a word), can be written in a unique manner in the form $r_1\vartheta(u)r_2$ with u an overlap-free word and r_1, r_2 are in $\{\varepsilon, a, b, aa, bb\}$. J. Cassaigne

[1] The symbol $\preccurlyeq$ has the meaning given by Bourbaki. In Landau notation, it is a big O.

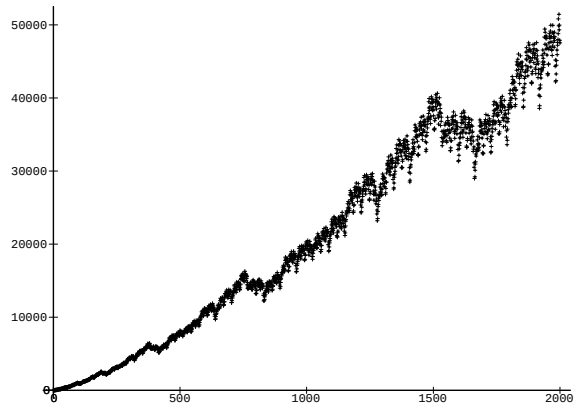

FIGURE 1. The number of overlap-free words presents a behaviour of polynomial type, but with heavy oscillations.

prefers to use a subtler decomposition, the advantage of which is to make independent both ends of the word.

He defines an action of the monoid $G = (E \times E)^*$, with $E = \{\delta, \iota, \kappa\}$, on the set $\{a, b\}^*$. The key idea is to modify the word $\vartheta(u)$ by deleting or inverting (in the sense that the inverse of a is b and conversely) letters at the ends of the word, in such a way that the resulting word is almost overlap-free if w is overlap-free and that every overlap-free word can be reached from S by this action. To be more precise, let U be the set of overlap-free words of length greater than 3, V be the set of words $xuxux$ with only an overlapping of only one letter x. J. Cassaigne proved the following result.

LEMMA 1. *Every word $w \in (U \cup V) \setminus S$ can be written $w = v.\gamma$ in a unique manner with $\gamma \in E \times E$ and $v \in U \cup V$. Moreover the length of v is less than the length of w.*

Here is why V is adequate. As a matter of fact, the word w may be overlap-free but v may be a minimal overlap.

Next, it is possible to iterate the process. Starting from an element $s \in S$, one applies a element $g \in G$ to obtain $z = s.g$. Technically, one has to remove some little words, so one uses only a subset Z of $S \times G$.

LEMMA 2. *Every word $w \in U \cup V$ has a unique decomposition $w = s.g$ with $(s, g) \in Z$.*

Automaton. The next crucial step is to consider the languages L and M of the words $(s, g) \in Z$ such that $s.g$ lies in U or V.

PROPOSITION 1. *The languages L and M are rational.*

Moreover, these languages are recognized by the same automaton, with only a change of the set of terminal states. The proof relies on the study of the overlappings of $s.g.\gamma$ with respect to γ and the overlappings of $s.g$. The states of the automaton are 2-tuples (i, j) with $i, j \in \{1, \ldots, 4\}$, which translate the prefix and suffix forms of the words. For example a word is of type $(2, 3)$ if it starts with *abb* or *baa* and ends with *abaa* or *babb*. The transitions of the automaton are labelled by the

elements of $E \times E$ and are expressed by $(i,j).(\gamma_1,\gamma_2) = (\varphi(i,\gamma_1),\varphi(j,\gamma_2))$. As we predicted the prefix and suffix of the words are disconnected. Roughly speaking, the automaton is made from three blocks: an initial block, which depends on the form of the words from S, and two blocks of the same structure, one which accepts the words in L and the other which accepts the words in M.

The automaton may be viewed in terms of matrices. Let $U_n(i,j)$, $V_n(i,j)$ be the number of words of length n in U and V respectively. Let δ, ι, κ be the matrix of the transformations δ, ι, κ,

$$\delta = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad \iota = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \kappa = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

PROPOSITION 2. *The matrices U_n and V_n satisfy the following recurrence,*

$$\begin{cases} V_{2n} &= 0, \\ V_{2n+1} &= \kappa {V_{n+1}}^t \delta + \delta {V_{n+1}}^t \kappa, \\ U_{2n} &= \iota {V_n}^t \iota + \delta {V_{n+1}}^t \delta + (\kappa+\iota) {U_n}^t (\kappa+\iota) + \delta {U_{n+1}}^t \delta, \\ U_{2n+1} &= \iota {V_{n+1}}^t \delta + \delta {V_{n+1}}^t \iota + (\kappa+\iota) U_{n+1}^{t\delta} + \delta {U_{n+1}}^t (\kappa+\iota), \end{cases}$$

for any integer n greater than 3.

The first few values,

$$V_4 = 0, \quad V_5 = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad U_4 = \begin{pmatrix} 2 & 0 & 2 & 0 \\ 0 & 2 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}, \quad U_5 = \begin{pmatrix} 0 & 2 & 2 & 0 \\ 2 & 0 & 0 & 2 \\ 2 & 0 & 2 & 0 \\ 0 & 2 & 0 & 0 \end{pmatrix},$$

enables computation of the matrices U_n and V_n, and hence

$$u_n = \sum_{1 \le i,j \le 4} U_n(i,j).$$

As a consequence of this recurrence, we have the following assertion.

THEOREM 1. *The sequence (u_n) of the number of overlap-free words is 2-regular.*

Recall the notion of B-regularity due to Allouche and Shallit [1].

2. Asymptotic behaviour

As a 2-regular sequence, u_n grows polynomially and it is natural to search for the best numbers α and β such that

$$n^{\alpha-\varepsilon} \preccurlyeq u_n \preccurlyeq n^{\beta+\varepsilon},$$

for any positive ε. Using integers the binary expansion of which has a given pattern (namely the numbers 7.2^k and $(22.4^k - 1)/3$), J. Cassaigne gives some possible values for α and β but not the best ones. He obtains, with the previous results of Kobayashi, the following inequalities for the best α and β

$$1.155 < \alpha < 1.276 < 1.332 < \beta < 1.587.$$

The sums

$$s_n = \sum_{k=0}^{n-1} u_k$$

are simpler to study because they give an increasing sequence; consideration of the integers $n = 7.2^k$ suffices to obtain the behaviour of s_n.

THEOREM 2. *The sequence s_n grows like $n^{\log_2 \zeta}$, where*

$$\zeta = \sqrt{3} + \frac{3 + \sqrt{5 + 4\sqrt{3}}}{2}.$$

As a matter of fact, the number ζ is the largest eigenvalue of a linear representation of the 2-regular sequence s_n. The result is not surprising because this is the situation in all the known cases [6], although there is no proof in the general case.

Bibliography

[1] Allouche (J.-P.) and Shallit (J.). – The ring of k-regular sequences. *Theoretical Computer Science*, vol. 98, 1992, pp. 163–197.

[2] Brlek (S.). – Enumeration of factors in the Thue-Morse word. *Discrete Applied Mathematics*, 1989, pp. 83–96.

[3] Carpi (Arturo). – Overlap-free words and finite automata. *Theoretical Computer Science*, vol. 115, 1993, pp. 243–260.

[4] Cassaigne (Julien). – Counting overlap-free binary words. In Enjalbert (Patrice), Finkel (A.), and Wagner (Klaus W.) (editors), *STACS '93. Lecture Notes in Computer Science*, vol. 665, pp. 216–225. – Würzburg, 1993.

[5] de Luca (A.) and Varricchio (S.). – Some combinatorial properties of the Thue-Morse sequence and a problem in semigroups. *Theoretical Computer Science*, vol. 63, 1989, pp. 333–348.

[6] Flajolet (Philippe), Grabner (Peter), Kirschenhofer (Peter), Prodinger (Helmut), and Tichy (Robert). – Mellin transforms and asymptotics: Digital sums. *Theoretical Computer Science*, vol. 123, n° 2, 1994, pp. 291–314.

[7] Kobayashi (Y.). – Enumeration of irreducible binary words. *Discrete Applied Mathematics*, vol. 20, 1988, pp. 221–232.

[8] Lothaire (M.). – *Combinatorics on Words*. – Addision–Wesley, 1983, *Encyclopedia of Mathematics and its Applications*, vol. 17.

[9] Restivo (A.) and Salemi (S.), Nivat (M.) and Perrin (D.) (editors). – *Overlap-free words on to symbols*, pp. 198–206. – Springer–Verlag, 1985, *Lectures Notes in Computer Science*, vol. 192.

[10] Thue (A.). – Über unendliche Zeichenreihen. *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl.*, n° 7, 1906, pp. 1–67. – Also in [11].

[11] Thue (A.). – *Selected Mathematical Papers*. – Universitetforlaget, Oslo, 1977. Edited by T. Nagell, A. Selberg, S. Selberg and K. Thalberg.