Lucian Ilie | Solomon Marcus | Ion Petre

Periodic and Sturmian languages

Turku Centre *for* Computer Science

TUCS Technical Report
No 629, May 2005

Periodic and Sturmian languages

LUCIAN ILIE
Department of Computer Science, University of Western Ontario
N6A 5B7, London, ON, CANADA
ilie@csd.uwo.ca
SOLOMON MARCUS
Romanian Academy, Mathematics
Calea Victoriei 125, Bucharest, ROMANIA
solomon.marcus@imar.ro
ION PETRE
Department of Computer Science, Åbo Akademi University,
Turku Center for Computer Science,
FIN-20520 Turku, Finland,
email: ipetre@abo.fi

TUCS Technical Report
No 629, May 2005

Abstract

Counting the number of distinct factors in the words of a language gives a measure of complexity for that language similar to the factor-complexity of infinite words. Similarly as for infinite words, we prove that this complexity function $f(n)$ is either bounded or $f(n) \geq n+1$. We call languages with bounded complexity *periodic* and languages with complexity $f(n) = n + 1$ *Sturmian*. We describe the structure of periodic languages and characterize the Sturmian languages as the sets of factors of (one- or two-way) infinite Sturmian words.

Keywords: infinite words, languages, factors, periodic, Sturmian

MSC: 68R15, 68Q45

TUCS Laboratory
Computational Biomodelling

1 Introduction

A function can be associated in a natural way to an infinite word by counting the number of factors of the same length. Fundamental results concerning this function and the implications on the structure of the underlying infinite word were proved already by Morse and Hedlund [14], Coven and Hedlund [4]. The most interesting cases are those corresponding to very low complexity of the above mentioned function, that is, bounded or marginally unbounded.

On the other hand, a similar function can be considered for languages of finite words. Already Berstel [1] considered the notion of the *population* function of a language L which associates, to every n, the number of words of length *at most* n in L. The notion of the number of words of the same length is certainly very basic one in language theory and it has been intensively studied already in [8, 9, 10]. Many results were discovered (or rediscovered) later in [6, 7, 13, 15, 18], to quote a few; [7] gives a good account of the history of the most important results. The same problem was also investigated for L-systems; see [16].

The above mentioned results were concerned only with special classes of languages, such as regular, context-free, D0L, etc. In turn, we consider here unrestricted languages but, in connection with infinite words, we count the number of factors of the same length. For this reason, we shall work with *factorial* languages, that is, languages which contain all factors of their words. Therefore, counting factors or words of the same length will make no difference. As it turns out, counting the number of factors of a given language gives enough information about the structure of the language, without the need of extra information about the language, such as its position in the Chomsky hierarchy.

For a (finite, one-way, or two-way infinite) word α, denote $f_n(\alpha)$ the number of factors of α of length n. As proved in [4], if α is right-infinite, then either $f_n(\alpha)$ is bounded, or $f_n(\alpha) \geq n + 1$, for all n. Right-infinite words α such that $f_n(\alpha)$ is bounded are called *periodic* (or often, *ultimately periodic*). Right-infinite words α with $f_n(\alpha) = n + 1$ for all $n \geq 0$, are called *Sturmian*, see [2]. Both these cases have been studied extensively in the literature, see, e.g. [11]. Two-way infinite words with the same property will be also called Sturmian for uniformity. They have been characterized by Coven and Hedlund [4].

We prove first a gap theorem, showing that the function f we investigate can be either bounded or at least linear. In the bounded case we obtain a class of languages, which we call *periodic*, for which all words have a short period except possibly for short prefixes and suffixes. We give a characterization of such languages similar with the one of Coven and Hedlund [4] for periodic words.

We consider then the marginally unbounded case, that is, when the number of words of the same length is always one greater than the length. We call such languages *Sturmian* in analogy with the terminology used for words. We characterize Sturmian languages as precisely the sets of factors of Sturmian (one- or two-way) infinite words, a rather unexpected result which reenforces the strength of the Sturmian property.

Several research directions are proposed. The connection between periodic and Sturmian infinite words and the corresponding periodic and Sturmian languages, respectively, is shown to be very strong. It seems interesting to investigate languages of higher complexity and see what is the relation with the infinite words.

2 Definitions

For basic notions and results on words we refer to [3], [11], [12] and for languages we refer to [5, 17].

An *alphabet* A is a finite nonempty set; A^* denotes the set of all finite words over A; ε is the empty word. The number of letters in A is denoted $|A|$. For a word $w \in A^*$, the *length* of w, which represents the number of letters of w, is denoted $|w|$. The set of words of length n (at most n) over A is denoted A^n ($A^{\leq n}$, resp.). A *language* is a subset of A^*. If $w = xyz$, for some words $w, x, y, z \in A^*$, then x is a *prefix* of w, y is a *factor* of w, and z is a *suffix* of w. For a language L, we shall denote by $\mathsf{fact}(L)$ the set of all factors of words in L and $\mathsf{fact}_n(L) = \mathsf{fact}(L) \cap A^n$; L is called *factorial* if $L = \mathsf{fact}(L)$. We make the convention that all languages considered throughout the paper are factorial unless otherwise specified. We also denote $\mathsf{pref}(L)$ the set of all prefixes of words in L.

A right-infinite word α is a function $\alpha : \mathbb{N} \to A$, where $\mathbb{N}$ is the set of positive integers. We also write $\alpha = \alpha_1\alpha_2 \ldots$, with $\alpha_i \in A$, for all $i \geq 1$. Any finite word $\alpha_{i,j} = \alpha_i\alpha_{i+1} \ldots \alpha_{i+j}$, $i \geq 1$, $j \geq 0$ is a *factor* of α; if $i = 1$, then $\alpha_{i,j}$ is also called a *prefix* of α. We denote the set of all finite factors of α by $\mathsf{fact}(\alpha)$. The notations $\mathsf{pref}(L)$ and $\mathsf{pref}(\alpha)$ are similarly defined for prefixes instead of factors. A bi-infinite (two-way infinite) word β is a function $\beta : \mathbb{Z} \to A$, where $\mathbb{Z}$ is the set of integers. We also write $\beta = \ldots \beta_{-1}\beta_0\beta_1 \ldots$, for all $i \in \mathbb{Z}$. Any finite word $\beta_i\beta_{i+1} \ldots \beta_{i+j}$, $i \in \mathbb{Z}$, $j \geq 0$ is a *factor* of β; we denote $\mathsf{fact}(\beta)$ the set of all factors of β.

For a language L and a non-negative integer n, we denote by $f_n(L)$ the number of words of length n in $\mathsf{fact}(L)$. Similarly, for a (right- or bi-) infinite word α, $f_n(\alpha)$ is the number of factors of length n in α.

An infinite word (one- or two-way) α is called *Sturmian* if, for any $n \geq 0$, $f_n(\alpha) = n + 1$.

The *factor graph* $G_n(L)$ of order n associated with a language L has the set of vertices $\mathsf{fact}_n(L)$ and the set of edges $\{(aw, wb) \mid a, b \in A, awb \in \mathsf{fact}_{n+1}(L)\}$.

3 Periodic languages

Coven and Hedlund [4] proved the following characterizations for the ultimate periodicity of a right-infinite word α using the function $f_n(\alpha)$.

Theorem 1 (Coven and Hedlund [4]). *Let α be a right-infinite word over A. The following assertions are equivalent:*

(i) $\alpha = uv^\omega$, *for some* $u \in A^*$, $v \in A^+$;

(ii) $f_n(\alpha)$ *is bounded;*

(iii) $f_n(\alpha) < n + k - 1$, *for some* $n \geq 1$, *where* k *is the number of letters in* α*;*

(iv) $f_n(\alpha) = f_{n+1}(\alpha)$, *for some* $n \geq 0$.

We shall find similar results for languages; the function $f_n(\alpha)$ is replaced by $f_n(L)$. The function $f_n(L)$ seems more complicated than the one for infinite words, see the examples below.

Example 1. (i) For any finite language L_1, $f_n(L_1) = 0$, for all sufficiently large n.

(ii) If $L_2 = ab^*$, then $f_n(L_2) = 2$, for all $n \geq 1$.

(iii) If $L_3 = A^{\leq k}$, for some $k \geq 1$, then $f_n(L_3) = |A|^n$, for all $n \leq k$ and $f_n(L_3) = 0$, for all $n > k$. Thus, the function f_n may not be uniform as in the case of infinite words.

(iv) If $L_4 = a^*b^*$, then $f_n(L_4) = n + 1$, for all $n \geq 0$.

(v) If $L_5 = a^*b^* \cup c^*$, then $f_n(L_5) = n + 2$, for all $n \geq 0$.

(vi) If $L_6 = a^*b^* \cup a^*c^*$, then $f_n(L_6) = 2n + 1$, for all $n \geq 0$.

(vii) If $L_7 = a(bb)^*c$, then $f_{2n+2}(L_7) = 4$ and $f_{2n+1}(L_7) = 3$, for all $n \geq 0$.

We shall need the following technical lemma in the proof of the main result of this section.

Lemma 2. *Let $z \in A^*$ and $n \geq 1$ such that the suffix of z of length $n - 1$ appears at least twice in z and $f_n(z) < n + 1$. Then $z = uv^s$ for some $u, v \in A^*$ with $|uv| \leq 3n$.*

Proof. The claim follows trivially if z contains a single letter. Assume then z contains at least two different letters. Any suffix of length $n - 1$ or less of z appears at least twice in z. Therefore, any factor of length $i \leq n - 1$ of z can be extended to the right to a factor of length $i + 1$ of z. Also, different factors of length i give different factors of length $i + 1$. Thus, $f_n(z) \geq f_{n-1}(z) \geq \cdots \geq f_1(z)$. If all these inequalities were strict, since $f_1(z) \geq 2$, we would obtain $f_n(z) \geq n + 1$, a contradiction. Therefore, there is $p, 1 \leq p \leq n - 1$, such that $f_p(z) = f_{p+1}(z)$. Consequently, in the graph $G_p(z)$, each vertex has exactly one outgoing edge and so, the strongly connected components of $G_p(L)$ are simple circuits. Thus, $z = uv^s$ where $|u| \leq p + f_p(z) \leq 2n$ and $|v| \leq f_p(z) \leq n$. □ □

We characterize now the languages with bounded complexity. For $k \geq 1$, we say that a language L is *k-periodic* if for any word $w \in L$, w has a factor of length at least $|w| - k$ with a period at most k. A language is *periodic* if it is k-periodic for some $k \geq 1$. In other words, a language is periodic if all its words have a small period, except possibly for a short prefix and suffix.

Theorem 3. *For a factorial language L over A the following assertions are equivalent:*

(i) L is periodic;

(i′) $L \subseteq \bigcup_{i=1}^{n} x_i y_i^* z_i$, for $n \geq 1$, $x_i, y_i, z_i \in A^*$, $|x_i z_i| \leq k, |y_i| \leq k$, $1 \leq i \leq n$;

(ii) $f_n(L)$ is bounded;

(iii) $f_{n_0}(L) < n_0 + 1$, for some $n_0 \geq 1$.

Proof. The equivalence between (i) and (i') is straightforward.

$(i') \Rightarrow (ii)$. We have that $f_n(xy^*z) \leq |x| + 2|y| + |z|$ is bounded and so is $f_n(L)$.

$(ii) \Rightarrow (i')$. Assume $f_n(L) \leq k$, for all n. Assume L infinite and consider an arbitrary word $w \in L$ such that $|w| \geq k+1$. Consider the factors of length $|w|-k$ of w. There are $k + 1$ such factors and so, using the bound on $f_n(L)$, there is a factor of w, say t, occurring twice in w. We have therefore $w = r_1 t r_2 = s_1 t s_2$, for some $r_i, s_i \in A^*$ with $|r_1 r_2| = |s_1 s_2| = k$. We may assume $|r_1| < |s_1|$. Then, there are $r, s \in A^*$ such that $s_1 = r_1 r$ and $r_2 = s s_2$, hence $rt = ts$. This equation gives that there are $x, y \in A^*$ such that $r = xy, s = yx, t \in (xy)^*x$. Consequently, $w \in r_1(xy)^+ x s_2$. Notice that $|xy| \leq k$ and $|r_1 x s_2| \leq |r_1 r_2| = k$.

As w has been arbitrarily chosen among the words of length at least $k+1$ in L, we have (i').

$(ii) \Rightarrow (iii)$. Obvious.

$(iii) \Rightarrow (ii)$. Assume that $f_n(L)$ is unbounded and consider m such that $f_m(L) > |A|^{4n_0}$. Consider an arbitrary word $z \in \mathsf{fact}_m(L)$. At most $f_{n_0}(L)$ factors of length n_0 of z can appear only once in z. Thus, there must be a prefix t of z of length at least $m - f_{n_0}(L)$ whose suffix of length n_0 appears at least twice in t. Applying Lemma 2 to t gives that $t = uv^s$ with $|uv| \le 3n_0$. Thus, any word z in $\mathsf{fact}_m(L)$ can be written as $z = uv^s w$ with $|uv| \le 3n_0$, $|w| \le f_{n_0}(L) \le n_0$. Hence $f_m(L) \le |A|^{4n_0}$, a contradiction. □ □

Remark 1. (a) We can show that (ii) implies L can be written as in (i') with $|xz| \le k-1$, $|y| \le k$. We have, using the notations in the proof of Theorem 3, $|r_1 x s_2| \le k$. If $|r_1 x s_2| = k$, then $|y| = 0$, hence y is the empty word. Therefore $w \in r_1 x^* s_2$ and $|r_1 s_2| < k$ since x cannot be empty. We obtained then the improvement that L is included in finite union of sets of the form xy^*z, with $|xz| \le k-1$ and $|y| \le k$. Moreover, these upper bounds on the lengths of the words $x_i z_i$ and y_i are optimal as shown by the language $L = (a^{k-1}b)^* a^{k-1}$.

(b) The condition in (iii) is tight. Consider $L = a^*b^*$; $f_n(L) = n+1$ and therefore unbounded. Also, we cannot replace condition (iii) by $f_n(L) < n + |A| - 1$ (similar with the corresponding one for infinite words). For instance, $f_n(a^*b^* + c) = n + 1 < n + |\{a, b, c\}| - 1$, for any $n \ge 2$, but is still unbounded.

(c) We have no condition such as (iv) in Theorem 1. For the precise formulation from words, we have the counterexample $L = a(bb)^*c$ for which $f_{2k}(L) = 4$, $f_{2k+1}(L) = 3$. On the other hand, $f_n(L)$ bounded obviously implies there are n and c such that $f_n(L) = f_{n+c}(L)$. We list as an open problem whether this property is sufficient for L to be bounded.

4 Sturmian languages

The main result of this section is the characterization of Sturmian languages. Clearly, for any infinite Sturmian word, its set of factors is a Sturmian language and we naturally ask whether there are some other Sturmian languages except for these – Theorem 4 gives a negative answer.

There is a huge literature about one-way infinite Sturmian words; [2] contains a brief survey of some of the most important results. The situation is simpler for two-way infinite Sturmian words. Coven and Hedlund [4] proved the two-way Sturmian words are precisely all words of the form

(i) ${}^{\omega}01^{\omega}$ or

(ii) $\tilde{y}zx$ with $y = (01p)^{\omega}$, $x = (10q)^{\omega}$, where $\tilde{y}$ is the reversal of y, p, q are palindromes and $z \notin 0^* \cup 1^*$ is a central word which has periods $|p|$ and $|q|$; see [2, problem 2.1.1].

They also proved that, as in the case of one-way infinite words, there is a gap between constant and Sturmian complexity for two-way infinite words.

We can prove now the main result of this section.

Theorem 4. *A factorial language is Sturmian if and only if it is the set of all finite factors of an infinite Sturmian word.*

Proof. We need to show only that the condition is necessary. Assume then L is a Sturmian language and consider the set

$$\mathrm{Inf}(L) = \{\alpha \in \{a, b\}^{\omega} \mid \mathsf{pref}(\alpha) \subseteq \mathsf{pref}(L)\}.$$

(Clearly, the alphabet of L must contain exactly two letters.) As L is Sturmian, it must be infinite and then König's lemma says that $\mathrm{Inf}(L)$ is nonempty. Also, for any $\alpha \in \mathrm{Inf}(L)$, we have also $\mathsf{fact}(\alpha) \subseteq \mathsf{fact}(L) = L$.

Since L is Sturmian, any $\alpha \in \mathrm{Inf}(L)$ satisfies $f_n(\alpha) \le n+1$, for all $n \ge 0$. If there is one such α with $f_n(\alpha) = n+1$, for all $n \ge 0$, then α is Sturmian and $L = \mathsf{fact}(\alpha)$, as claimed.

Assume this is not the case for any word of $\mathrm{Inf}(L)$. This means, for any $\alpha \in \mathrm{Inf}(L)$, there is $n \ge 1$ such that $f_n(\alpha) < n+1$. By Theorem 1, any $\alpha \in \mathrm{Inf}(L)$ is then periodic and $f_n(\alpha)$ is bounded. Therefore, there must be infinitely many different infinite words in $\mathrm{Inf}(L)$. Indeed, for any finitely many such infinite words, L still contains infinitely many words which do not appear as factors in any of those and so we can construct more infinite words in $\mathrm{Inf}(L)$ by König's lemma.

On the other hand, all words in $\mathrm{Inf}(L)$ are periodic and therefore $\mathrm{Inf}(L)$ is countably infinite; put $\mathrm{Inf}(L) = \{u_i v_i^\omega \mid i \ge 1\}$, where, for any $i \ge 1$, v_i is primitive and moreover, u_i and v_i have no nontrivial common suffix. Hence, $f_{|v_i|}(u_i v_i^\omega) \ge |v_i|$ and there is at most one factor of length $|v_i|$ in any $u_j v_j^\omega$, $j \ne i$, which does not appear in $u_i v_i^\omega$.

We prove first that there cannot be infinitely many different words v_i. Assume there are infinitely many. Therefore, we can find three of those, say v_i, v_j, v_k, such that $2 \le |v_i| < |v_j|$ and $|v_k| \ge 2(|v_i| + |v_j|)$. As argued above, there is a factor of v_k^ω of length $|v_k|$ which is also a factor of v_i^ω and similarly for v_j^ω. These two factors overlap at least half their length in v_k^ω — this gives a word of length at least $|v_i|+|v_j|$ which has periods $|v_i|$ and $|v_j|$. Fine and Wilf's periodicity lemma (see [12]) implies that v_j is not primitive, a contradiction.

Thus, there are only finitely many v_i and so there is one, denoted v in the following, such that the set $V = \{u_i \in A^* \mid u_i v^\omega \in \mathrm{Inf}(L)\}$ is infinite.

For any i such that $u_i \in V$ is nonempty, u_i and v have no nonempty common suffix, they end with different letters; say a for v and b for u_i. As there are infinitely many, we can construct a left-infinite word α (by König's lemma) such that any suffix of α is a suffix of some $u_i \in V$.

Now, for any sufficiently large n, we have that $f_n(\alpha b v^\omega) \ge n+1$. Indeed, this follows from the result of Coven and Hedlund [4] stating that bi-infinite words cannot have any complexity between bounded and Sturmian. Since for any $n \ge |v|+1$, $f_n(\alpha b v^\omega) \ge n - |v| + 1$, it is unbounded and therefore the complexity must be at least Sturmian.

On the other hand, since $\mathsf{fact}(\alpha b v^\omega) \subseteq \mathsf{fact}(L)$ and L is Sturmian, it follows that $f_n(\alpha b v^\omega) = n+1$. Therefore, $\alpha b v^\omega$ is Sturmian and $L = \mathsf{fact}(\alpha b v^\omega)$, concluding the proof of the theorem. □ □

5 Conclusions and open problems

We can easily construct factorial languages from infinite words by taking the set of all finite factors. Having a factorial language, we can construct the right-infinite words in $\mathrm{Inf}(L)$ or the corresponding sets for left- and bi-infinite. It is interesting to notice that with these constructions, the notions of periodicity and Sturmian for infinite words and languages are very closely related. Indeed, from periodic and Sturmian infinite words we obtain periodic and Sturmian languages, resp. Conversely, from periodic languages we construct periodic words, while from Sturmian languages we construct either right- or bi-infinite Sturmian words.

We list here only a few open problems suggested by our results.

Problem 1. Does $f_n(L) = f_{n+c}(L)$ imply $f_n(L)$ bounded?

Problem 2. $L = a(bb)^*c$ shows that, in the case $f_n(L)$ is bounded, it might never become monotonic. Is this possible in the unbounded case?

Problem 3. Give other interesting characterizations for $f_n(L)$ to be bounded in terms of properties of $f_n(L)$.

Problem 4. Characterize the function $f_n(L)$.

Problem 5. Characterize the languages with low unbounded complexity; e.g., $f_n(L) = n + c$ or linear in n. Some examples here are $L_1 = a^*b_1b_2\ldots b_ca^*$ with $f_n(L_1) = n+c$, for all $n \geq 1$, and $L_2 = a^*ba^* + a^*bc^*$ with $f_n(L_2) = 2n+1$, for all $n \geq 1$.

Problem 6. It is interesting to see to what extent complexities higher than Sturmian are preserved when moving from languages to infinite words and viceversa.

Acknowledgements

Useful comments from the anonymous referee are gratefully acknowledged. Lucian Ilie was supported in part by NSERC. Ion Petre was supported by Academy of Finland, project 203667.

References

[1] Berstel, J., Sur la densité asymptotique de langages formels, in M. Nivat, ed., *Automata, Languages, and Programming* (North-Holland, 1972) 345 – 358.

[2] Berstel, J., and P. Séébold, Sturmian Words, in [12], pp 45 – 110.

[3] Choffrut, C., and J. Karhumäki, Combinatorics on Words. In G. Rozenberg, A. Salomaa (eds.), *Handbook of Formal Languages*, vol. 1: 329-438, Springer-Verlag, 1997.

[4] Coven, E.M, and G.A. Hedlund, Sequences with minimal block growth, *Math. Systems Theory* **7** (1973) 138 – 153.

[5] Hopcroft, J.E., and J.D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, Reading, Mass., 1979.

[6] Ilie, L., On a conjecture about slender context-free languages, *Theoret. Comput. Sci.* **132** (1994) 427 – 434.

[7] Ilie, L., G. Rozenberg, and A. Salomaa, A characterization of poly-slender context-free languages, *Theoret Inform. Appl. (RAIRO)* **34** (2000) 77 - 86.

[8] Kunze, M., H.J. Shyr, and G. Thierrin, h-bounded and semidiscrete languages, *Inform. and Control* **51** (1981) 147 – 187.

[9] Latteux, M., and G. Thierrin, Semidiscrete context-free languages, *Internat. J. Comput. Math.* **14** (1983) 3 – 18.

[10] Latteux, M., and G. Thierrin, On bounded context-free languages, *Elektron. Informtionsverarb. Kybernet.* **20** (1984) 3 – 8.

[11] Lothaire, M., *Combinatorics on Words*, Addison-Wesley, Reading, MA., 1983.

[12] Lothaire, M., *Algebraic Combinatorics on Words*, Cambridge Univ. Press, 2002.

[13] Marcus, S., On the lengths of words, in: J. Karhumäki, H. Maurer, Gh. Păun, G. Rozenberg, *Jewels are Forever*, Springer-Verlag, Berlin, 1999, 194 – 203.

[14] Morse, M., and G.A. Hedlund, Symbolic dynamics II. Sturmian trajectories, *Amer. J. Math.* **61** (1940) 1 – 42.

[15] Păun, Gh., and A. Salomaa, Thin and slender languages, *Discrete Appl. Math.* **61** (1995) 257 – 270.

[16] Rozenberg, G., On subwords of formal languages, *Proc. of Fundamentals of Computation Theory (Szeged, 1981)*, Lecture Notes in Comput. Sci. **117**, Springer, Berlin-New York, 1981, 328 - 333

[17] Salomaa, A., *Formal Languages*, Academic Press, 1973.

[18] Shallit, J., Numeration systems, linear recurrences, and regular sets, *Inform. and Comput.* **113** (1994) 331 – 347.

Turku

Centre *for*

Computer

Science

Lemminkäisenkatu 14 A, 20520 Turku, Finland | www.tucs.fi

University of Turku

- Department of Information Technology
- Department of Mathematical Sciences

Åbo Akademi University

- Department of Computer Science
- Institute for Advanced Management Systems Research

Turku School of Economics and Business Administration

- Institute of Information Systems Sciences

ISBN 952-12-1443-0
ISSN 1239-1891