RAIRO-Inf. Theor. Appl. **39** (2005) 641-650 DOI: 10.1051/ita:2005033

# LANGUAGES OF FINITE WORDS OCCURRING INFINITELY MANY TIMES IN AN INFINITE WORD

## KLAUS THOMSEN<sup>1</sup>

**Abstract.** We give necessary and sufficient conditions for a language to be the language of finite words that occur infinitely many times in an infinite word.

Mathematics Subject Classification. 37B10, 68R15.

## 1. INTRODUCTION

In dynamical systems there is a notion of  $\omega$ -limits of a point or set of points, which is fundamental for several important parts of the theory. The omega limit set  $\omega(x)$  of a point x in a dynamical system consists of the points all of whose neighbourhoods are visited infinitely many times under the iteration of the map, starting at the point x. The present work arose from the need to determine the structure of the  $\omega$ -limit set of a point in a one-sided shift space. Since  $\omega(x)$  is always a closed invariant subset of the dynamical system, the  $\omega$ -limit set of a point of a one-sided shift space is itself a shift space, and can therefore equally well be considered as a language. The well-known bijective correspondance between shift spaces and a particular class of languages makes it possible, and also helpful to consider the question in its language-theoretic guises. It then takes the following form: which languages arise as the language of the finite words occurring infinitely many times in an infinite word? In this form the question was addressed in [6], together other related questions. We will show here that the languages in question are those that are *infinite*, factorial, and chain transitive. "Infinite" means that the language must contain infinitely many words, and "factorial" is the familiar condition that the language must be stable under the passage to subwords, while the third condition, "chain transitivity" may be new in the language setting, although its dynamical systems-equivalent is well established. Given these three necessary and sufficient conditions, it is easy to see that many regular languages

 $\bigodot$  EDP Sciences 2005

Keywords and phrases. Formal languages, infinite words, infinite occurrence.

<sup>&</sup>lt;sup>1</sup> Institut for matematiske fag, Ny Munkegade, 8000 Aarhus C, Denmark; matkt@imf.au.dk

#### K. THOMSEN

can be realized as the language of the finite words occurring infinitely many times in an infinite word, while just as many can not. We show that there is a finite algorithm from which it is easy to determine if a given finite automaton, with all states initial and final, will give rise to a language of finite words occurring infinitely many times in an infinite word, or not. In this sense, the question is certainly decidable for regular languages. Finally, we answer onother question about languages arising from infinite words which was explicitly stated as an open problem in [6]. Other open problems from [6] have been solved in [2] and [4].

#### 2. Terminology and basic definitions

Let A be a finite alphabet. The set of finite words in A (including the empty word) will be denoted by  $A^*$ , and the length of a word  $u \in A^*$  will be denoted by |u|. An infinite word in the alphabet A is formally a function  $\mathbb{N} \to A$ , and is presented as a right-infinite string of letters from the alphabet. The set of infinite words in A will be denoted by  $A^{\mathbb{N}}$ . The shift  $\sigma$  acts on  $A^{\mathbb{N}}$  in the usual way:

$$\sigma\left(a_1a_2a_3a_4\ldots\right) = a_2a_3a_4a_5\ldots$$

When  $u \in A^*$  and  $a = a_1 a_2 \cdots \in A^* \cup A^{\mathbb{N}}$ , we write  $u \subseteq a$  to mean that  $u = a_{i+1}a_{i+2} \ldots a_{i+|u|}$  for some *i*, and we let  $p(a) \subseteq A^*$  be the set of prefixes of *a*, and  $s(a) \subseteq A^*$  the set of suffixes of *a*, provided of course that  $a \in A^*$ . When  $a \in A^{\mathbb{N}}$ , we let  $L_{\infty}(a)$  denote the language consisting of the finite words occurring infinitely many times in *a*, *i.e.* 

$$L_{\infty}(a) = \{ u \in A^* : u = a_{i+1}a_{i+2}a_{i+3} \dots a_{i+|u|} \text{ for infinitely many } i \}.$$

When investigating the structure of such a language, it is natural to consider two cases separately. An element  $a \in A^{\mathbb{N}}$  is *recurrent* when every finite word which appears in a must appear infinitely many times, *i.e.* when  $L_{\infty}(a) = \{u \in A^* : u \subseteq a\}$ . We will say that a is *eventually recurrent* when  $\sigma^k(a)$  is recurrent for some  $k \in \mathbb{N}$ . A language L in the alphabet A is *transitive* when  $w_1, w_2 \in L \Rightarrow \exists u \in A^* : w_1 u w_2 \in L$ . We call L factorial when  $u \subseteq w \in L \Rightarrow u \in L$ .

### 3. The language used to write a recurrent infinite word

**Proposition 3.1.** Let L be a non-empty language in the finite alphabet A. Then the following are equivalent:

- 1)  $L = L_{\infty}(a)$  for an infinite word  $a \in A^{\mathbb{N}}$  which is eventually recurrent.
- 2) L is factorial and transitive.
- 3)  $L = L_{\infty}(a)$  for an infinite word  $a \in A^{\mathbb{N}}$  which is recurrent.

*Proof.* 1)  $\Rightarrow$  2): it is obvious that L is factorial. Choose  $k \in \mathbb{N}$  such that  $\sigma^k(a)$  is recurrent. Let  $w_1, w_2 \in L$ . Since both words occur infinitely often in a, there

are occurrences of both words in  $\sigma^k(a)$  such that  $w_2$  is strictly to the right of  $w_1$ . It follows that there is a word  $u \in A^*$  such that  $w_1 u w_2 \subseteq \sigma^k(a)$ . Since  $\sigma^k(a)$  is recurrent,  $w_1 u w_2$  must occur infinitely often in  $\sigma^k(a)$ , and therefore also in a. Thus  $w_1 u w_2 \in L$ .

2)  $\Rightarrow$  3): let  $w_1, w_2, w_3, \ldots$  be any numbering of the non-empty words in L. Since L is transitive we can choose words  $a_i \in A^*$  such that  $w_1a_1w_2a_2w_3a_3\ldots a_{n-1}w_n \in L$  for all n. Then

$$x = w_1 a_1 w_2 a_2 w_3 a_3 \dots \in A^{\mathbb{N}}$$

is an infinite word in A, and  $L_{\infty}(x) \subseteq L$  because L is factorial. To see that  $L = L_{\infty}(x)$ , it suffices to consider  $i, j \in \mathbb{N}$  and show that there is a  $k \geq j$  such that  $w_i \subseteq w_k$ . To this end note that  $w_i \subseteq w_1 a_1 w_2 a_2 \dots a_{m-1} w_m \in L$ , when  $m = \max\{i, j\}$ , and that  $w_1 a_1 w_2 a_2 \dots a_{m-1} w_m = w_k$  for some k, which must be larger than j since  $|w_k| > |w_l|$  for all  $l \leq j$ . It remains only to show that x is recurrent, *i.e.* that every prefix of x occurs infinitely many times in x. But a prefix of x is contained in  $w_1 a_1 w_2 a_2 w_3 a_3 \dots a_{n-1} w_n$  for some n and must therefore be a word in L since L is factorial. Thus any prefix of x must equal  $w_i$  for some i, and each  $w_i$  occurs infinitely many times in x, as we have just seen.

 $(3) \Rightarrow 1)$  is trivial.

### 4. When the infinite word is not eventually recurrent

We turn to the case of infinite words that are not eventually recurrent. Let L be a factorial language in the finite alphabet A. Set

$$\mathsf{C}L = \{ w \in A^* : w \notin L \}.$$

An element  $u \in CL$  is minimal when  $w \in A^*$ ,  $w \subseteq u$ ,  $|w| < |u| \Rightarrow w \in L$ . The set of minimal words in CL will be denoted by MIN (CL). We say that L is of finite type when # MIN (CL)  $< \infty$ . Let  $w, v \in L$ . A tile from w to v in L is a word  $u \in L$  such that  $w \in p(u)$  and  $v \in s(u)$ . Such a tile will be denoted symbolically by

$$w \xrightarrow{u} v,$$

or simply by  $w \to v$ , when we don't need to specify the word u. A path from w to v in L is a sequence of tiles

$$w_0 \xrightarrow{u_1} w_1 \xrightarrow{u_2} w_2 \xrightarrow{u_3} \dots \xrightarrow{u_N} w_{N+1}$$

such that  $w = w_0, v = w_{N+1}$ , and  $|w_i| \ge \min\{|w|, |v|\}$  for all i = 1, 2, ..., N. We say that L is *chain transitive* when every pair of words in L can be connected by a path.

**Lemma 4.1.** Let L be an infinite language which is chain transitive. Then L is prolongable, in the sense that for every word  $a_1a_2...a_n \in L$  there are letters  $a_0, a_{n+1} \in A$  such that  $a_0a_1a_2...a_na_{n+1} \in L$ .

*Proof.* Left to the reader.

**Lemma 4.2.** Assume that L is a factorial language in A. Let  $w, v \in L$  and assume that there is a path from w to v in L. Let y be a word in  $A^*$  such that  $w \in s(y)$ . It follows that there is a word  $x \in A^*$  such that  $y \in p(x), v \in s(x)$  and

$$\begin{array}{c} u \subseteq x, \\ u \notin y, \\ |u| < \min\{|w|, |v|\} \end{array} \right\} \Rightarrow u \in L.$$

$$(4.1)$$

*Proof.* Assume first that that  $w \xrightarrow{u} v$ . Write  $u = u_1 u_2 \dots u_n$  such that  $w = u_1 u_2 \dots u_{|w|}$  and  $v = u_{n-|v|+1} u_{n-|v|+2} \dots u_n$ . Then  $x = y u_{|w|+1} u_{|w|+2} \dots u_n \in A^*$  has the property described in (4.1). Proceed by induction.

**Proposition 4.3.** Let L be a language in the alphabet A. Then the following are equivalent:

- 1)  $L = L_{\infty}(x)$  for an infinite word  $x \in A^{\mathbb{N}}$  which is not eventually recurrent.
- 2) L is factorial, infinite, chain transitive and not of finite type.

*Proof.* 1)  $\Rightarrow$  2): *L* is obviously factorial, and it is easy to see that it is also infinite. (This follows also from Th. 9 of [6].) By deleting a finite prefix of *x* we may assume that  $a \in L_{\infty}(x)$  for all  $a \in A$ . Write  $x = x_0 x_1 x_2 \ldots$  We claim that there are sequences  $i_0 < i_1 < i_2 < \ldots$  in  $\mathbb{N}$  and  $s_0, s_1, s_2, \ldots$  in MIN (CL) such that

$$s_k = x_{i_k+1} x_{i_k+2} \dots x_{i_k+|s_k|}, \tag{4.2}$$

for all k, and

$$x_{i_{k-1}+2}x_{i_{k-1}+3}\dots x_{i_k+|s_k|-1} \in L$$
(4.3)

for all  $k \ge 1$ . We construct these two sequences by induction. So assume that we have found a string  $s_0, s_1, \ldots, s_n$  in MIN (CL) and  $i_0 < i_1 < \cdots < i_n$  in N such that (4.2) holds for all  $k \le n$  and (4.3) holds for all  $1 \le k \le n$ . Since the infinite word  $x_{i_n+2}x_{i_n+3}x_{i_n+4}\dots$  is not recurrent, there is an  $N > i_n + |s_n|$  such that  $x_{i_n+2}x_{i_n+3}\dots x_N \notin L_{\infty}(x)$ . Set

$$d = \min \left\{ j \in \mathbb{N} : j \ge i_n + 2 \text{ and } x_{i_n+2} x_{i_n+3} \dots x_j \notin L_{\infty}(x) \right\}.$$

Then  $i_n + 2 < d \leq N$ , and  $i_n + |s_n| < d$  since  $s_n \in MIN(CL)$ . Furthermore,  $x_{i_n+2}x_{i_n+3}\ldots x_{d-1} \in L$ . Set

$$i_{n+1} = \max\{j \le d : x_{j+1}x_{j+2} \dots x_d \notin L_{\infty}(x)\},\$$

and note that  $i_n + 1 \leq i_{n+1}$ . It follows that  $s_{n+1} = x_{i_{n+1}+1}x_{i_{n+1}+2}x_{i_{n+1}+3}\dots x_d \in MIN$  (CL). This completes the construction. Note that an  $s_i$  can only occur finitely many times in the sequence since otherwise it would be an element of  $L_{\infty}(x)$ . It follows that

$$\lim_{i \to \infty} |s_i| = \infty. \tag{4.4}$$

In particular, we conclude from this that  $\# \operatorname{MIN}(\operatorname{CL}) = \infty$ , *i.e.* L is not of finite type. Next we use the  $s_k$ 's to prove that L is chain transitive. Let w and v be non-empty words in L. Note that w can not contain any element from MIN (CL) because L is factorial. Since w occurs in x, there must be an  $r \in \mathbb{N}$  such that  $w \subseteq x_{i_{r-1}+2}x_{i_{r-1}+3}\ldots x_{i_r+|s_r|-1}$ ; specifically  $w = x_ax_{a+1}\ldots x_b$ , where  $i_{r-1}+2 \leq a \leq b \leq i_r+|s_r|-1$ . Furthermore, since w occurs infinitely many times in x, it follows from (4.4) that we can arrange that  $|s_j| \geq \min\{|w|, |v|\} + 3$  for all  $j \geq r$ . Similar considerations apply to v, so we see that there is a t > r such that  $v \subseteq x_{i_{t-1}+2}x_{i_{t-1}+3}\ldots x_{i_t+|s_t|-1}$ . It suffices therefore to show that for any non-empty word  $w'' \subseteq x_k x_{k+1} \ldots x_{i_{r+1}+|s_{r+1}|-1}$ , where  $k = \max\{a, i_r+2\}$ , with  $|w''| \geq \min\{|w|, |v|\}$ , there are tiles  $w \to w' \to w''$  such that  $|w'| \geq \min\{|w|, |v|\}$ , or a tile such that  $w \to w''$ . To this end, let  $w'' = x_k x_{k+1} \ldots x_{i_r+1+|s_{r+1}|-1}$ . If  $a < i_r + 2$ , set  $u_2 = x_{i_r+2}x_{i_r+3} \ldots x_{i_{r+1}+|s_{r+1}|-1}$ ,  $u_1 = x_a x_{a+1} \ldots x_{i_r+|s_r|-1}$  and  $w' = x_{i_r+2}x_{i_r+3} \ldots x_{i_r+|s_r|-1}$ . Then

$$w \xrightarrow{u_1} w' \xrightarrow{u_2} w''.$$

If  $a \ge i_r + 2$ , set  $u = x_a x_{a+1} \dots x_{i_{r+1}+|s_{r+1}|-1}$ , and note that  $w \xrightarrow{u} w''$  in this case. It follows that  $L = L_{\infty}(x)$  is chain transitive.

 $(2) \Rightarrow 1$ : since L is not of finite type, there is a sequence  $s_0, s_1, s_2, \ldots$  of distinct elements in MIN (CL), all of length at least 3. Let  $w_0, w_1, w_2, w_3, \ldots$  be a list of the non-empty words in L. For convenience we arrange the sequences such that  $|s_i| \leq |s_{i+1}|$  and  $|w_i| \leq |w_{i+1}|$  for all i. Note that  $\lim_{k\to\infty} |s_k| = \lim_{k\to\infty} |w_k| = \infty$ . For every word  $x = a_1 a_2 \ldots a_n \in A^*$  of length  $n \geq 3$ , let m(x), l(x) and r(x) denote the words  $m(x) = a_2 a_3 \ldots a_{n-1}, l(x) = a_1 a_2 \ldots a_{n-1}$  and  $r(x) = a_2 a_3 \ldots a_n$ , respectively. Then  $m(s_i), l(s_i), r(s_i) \in L$  for all i. Since L is chain transitive there is, for each i, a path from  $r(s_i)$  to  $w_{i+1}$  and from  $w_{i+1}$  to  $l(s_{i+1})$ . We use these paths to construct a sequence of words,  $a_1, a_2, a_3, \ldots \in A^*$  such that

a)  $a_i \in p(a_{i+1});$ b)  $s_i \in s(a_{2i});$ c)  $w_i \in s(a_{2i-1});$  and d)

$$\left. \begin{array}{l} u \subseteq a_{j}, \\ u \notin a_{2i}, \\ |u| < \min\{|m(s_{i-1})|, |w_{i-1}|\} \end{array} \right\} \Rightarrow u \in L,$$

for all  $i \ge 1$ . The construction proceeds by an induction started by taking  $a_1 = w_1$ . Assume that  $a_1, a_2, \ldots, a_{2n}$  have been constructed. By Lemma 4.2 there is word  $a_{2n+1} \in A^*$  such that  $a_{2n} \in p(a_{2n+1}), w_{n+1} \in s(a_{2n+1})$  and

$$\left.\begin{array}{l} u \subseteq a_{2n+1}, \\ u \nsubseteq a_{2n}, \\ |u| < \min\{|r(s_n)|, |w_{n+1}|\} \end{array}\right\} \Rightarrow u \in L.$$

Another application of Lemma 4.2 gives us a word  $y \in A^*$  such that  $a_{2n+1} \in p(y), l(s_{n+1}) \in s(y)$ , and

$$\left. \begin{array}{l} u \subseteq y, \\ u \not\subseteq a_{2n+1}, \\ |u| < \min\{|l(s_{n+1})|, |w_{n+1}|\} \end{array} \right\} \Rightarrow u \in L.$$

Set  $a_{2n+2} = yz$ , where  $z \in A$  is the last letter in  $s_{n+1}$ . Then  $a_{2n+1} \in p(a_{2n+2})$ ,  $s_{n+1} \in s(a_{2n+2})$ , and

$$\left.\begin{array}{l}u \subseteq a_{2n+2},\\ u \not\subseteq a_{2n},\\ |u| < \min\{|m\left(s_{n}\right)|, |w_{n}|\}\end{array}\right\} \Rightarrow u \in L.$$

Thus the desired sequence of  $a_i$ 's can be constructed by induction. It follows from condition a) that there is a unique infinite word  $x \in A^{\mathbb{N}}$  such that  $a_i$  is a prefix of x for all i. We claim that  $w_i \in L_{\infty}(x)$ . To see this, note that that  $w_i \subseteq w_j$  for infinitely many j because L is prolongable by Lemma 4.1. By construction  $w_j \subseteq x$  for all j, and we conclude therefore that  $w_i \in L_{\infty}(x)$ . Thus  $L \subseteq L_{\infty}(x)$ . It follows from d) that the reversed inclusion also holds. Indeed, if  $w \in L_{\infty}(x)$ , we can choose an  $i \in \mathbb{N}$  so large that  $\min \{|m(s_{i-1})|, |w_{i-1}|\} > |w|$ . Since  $w \in L_{\infty}(x)$ , there must be a j > 2i such that  $w \subseteq a_j$ , but  $w \not\subseteq a_{2i}$ . Hence  $w \in L$  by condition d).

Up to now we haven't used the presence of the  $s_i$ 's in x. They are needed now to show that x can not be eventually recurrent: For every  $k \in \mathbb{N}$  there is by construction an  $s_i$  in  $x_k x_{k+1} x_{k+2} \ldots$ , and hence L is not equal to  $\{u \in A^* : u \subseteq x_k x_{k+1} x_{k+2} \ldots\}$ , and x is therefore not eventually recurrent.  $\Box$ 

**Lemma 4.4.** Let L be a factorial and chain transitive language. Assume that L is of finite type. It follows that L is transitive.

*Proof.* Let  $k \in \mathbb{N}$  be larger than the maximal length of a word in MIN (LL). It follows then that every word w of L of length  $|w| \ge k$  is synchronizing, in the sense that the implication

$$\begin{array}{l} u, v \in A^*, \\ uw, wv \in L \end{array} \right\} \Rightarrow uwv \in L$$
 (4.5)

holds. Consider then any pair of words  $w_1, w_2 \in L$ . Since L is prolongable by Lemma 4.1 there are words  $w, v \in L$  such that  $\min\{|w|, |v|\} \geq k, w_1 \in p(w)$  and  $w_2 \in s(v)$ . Since L is chain transitive there is path from w to v with all tiles longer than k. It follows then from (4.5) that this path can put together to form a tile

646

from w to v. This shows that for any two words,  $w_1$  and  $w_2$ , in L, we can find a word  $u \in L$  such that  $w_1 \in p(u)$  and  $w_2 \in s(u)$ . Since L is also infinite, it must be transitive.

**Theorem 4.5.** A language L in the finite alphabet A is the language of the words occurring infinitely many times in an infinite word in A if and only if L is

- *infinite;*
- factorial; and
- chain transitive.

The infinite word can be chosen to be recurrent if and only if L is transitive.

*Proof.* Combine Proposition 3.1, Proposition 4.3 and Lemma 4.5.

Note that it follows from Propositions 3.1 and 4.3 that a language which is infinite, factorial and transitive, but not of finite type, can be realized as  $L_{\infty}(x)$ , both with an x which is recurrent and with an x which is not eventually recurrent.

In general it is not so easy to determine if a given language is chain transitive or not. For regular languages, however, it is quite easy, as we shall explain in the next section.

# 5. Which infinite and factorial regular languages are chain transitive?

A regular and factorial language L can be described as the words recognized by a finite automatum in which all states are initial and final, cf. Lemma 5 of [2]. If L is also prolongable, as it must be if it is chain transitive by Lemma 4.1, it is the language of a sofic shift space, [5], and L consists of the finite words obtained by "reading" the finite paths in a finite graph without sinks or sources, where the edges are labeled by the alphabet A, cf. [5]. The question we are looking at is therefore the following: given a finite graph without sinks or sources in which the edges are labeled by the letters of the alphabet A, how do we decide if the corresponding language is chain transitive?

Let G be a finite labeled graph without sinks or sources. Let  $C_1, C_2, \ldots, C_N$  denote the communication classes of vertices in G, cf. Section 4.4 in [5]. For each  $C_i$  we have an irreducible subgraph  $G_i$  of G whose vertices are the vertices in  $C_i$  and whose edges consist of the edges from G that go between vertices in  $C_i$ . Each  $G_i$  inherits the labeling from G and defines a sub-language  $L_i$  of the language L defined by G. Each  $L_i$  is infinite, factorial and transitive. Define an oriented graph  $\mathcal{G}$  whose vertices are  $1, 2, \ldots, N$ . There is an edge in  $\mathcal{G}$  from i to j if and only  $i \neq j$  and either  $\# (L_i \cap L_j) = \infty$  or there is a path in G from some vertex in  $C_i$  to a vertex in  $C_j$ . We call  $\mathcal{G}$  the graph of chain connected components of G.

**Lemma 5.1.** Let M be the square of the number of edges in G. Then  $\#(L_i \cap L_j) = \infty$  if and only if  $L_i \cap L_j$  contains a word of length M.

#### K. THOMSEN

Proof. Note that  $L_i \cap L_j$  consists of the words labeled by a path in the labeled graph  $H_{i,j}$  with vertices  $C_i \times C_j$ , where there is an edge labeled a from (e, f) to (e', f') if and only if there are edges  $e \to e'$  and  $f \to f'$  in  $G_i$  and  $G_j$ , respectively, both labeled a. It follows that any path in  $H_{i,j}$  with more that M edges must contain a loop.

**Proposition 5.2.** The language L determined by the labeled graph G is chain transitive if and only if the graph  $\mathcal{G}$  of chain connected components of G is irreducible (i.e. for  $i \neq j$  there is a path in  $\mathcal{G}$  from i to j).

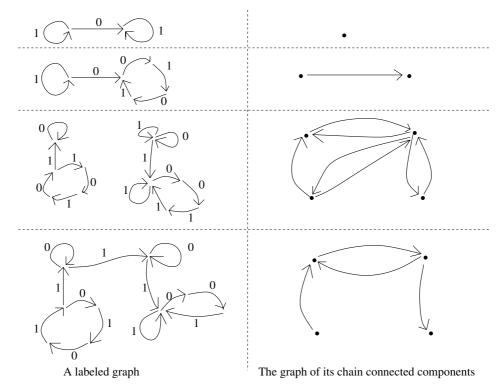
*Proof.* Assume first that  $\mathcal{G}$  is irreducible, and consider two words  $w_1, w_2 \in L$ . Let  $p_i$  be a path in G which label  $w_i, i = 1, 2$ . By prolonging  $p_1$  to the right and  $p_2$ to the left, we can obtain two new paths,  $p_3$  and  $p_4$ , respectively, such that the terminal vertex of  $p_3$  is an element of some  $C_i$  while the initial vertex of  $p_4$  is an element of some  $C_i$ . If i = j, we can connect the two vertices by a path in  $G_i$ and placed together with  $p_3$  and  $p_4$  we obtain a path in G which label a word  $u \in L$  such that  $w_1 \in p(u)$  and  $w_2 \in s(u)$ . Thus u is a tile from  $w_1$  and  $w_2$ . If  $i \neq j$ , there is by assumption a path in  $\mathcal{G}$  from  $C_i$  to  $C_j$ . Consider the first edge  $C_i \to C_{i'}$  in this path. If there is a path in G from some vertex in  $C_i$  to a vertex in  $C_{i'}$  we can proceed as above to obtain a tile u from  $w_1$  to a word  $w'_1 \in L_{i'}$ with  $|w'_1| \geq \min\{|w_1|, |w_2|\}$ . If instead  $\#(L_i \cap L_{i'}) = \infty$ ,  $L_i \cap L_{i'}$  must contain a word  $w'_1$  with  $|w'_1| \ge \min\{|w_1|, |w_2|\}$ . Let  $p_5$  be a path in  $G_i$  which label  $w'_1$ . Since  $G_i$  is irreducible we can connect  $p_3$  with  $p_5$  to obtain a tile from  $w_1$  to  $w'_1$ . We can therefore proceed through finitely many steps (one step for each of the edges in the path in  $\mathcal{G}$  connecting *i* to *j*), to get a path from  $w_1$  to a word  $v \in L_j$ such that  $|v| \ge \min\{|w_1|, |w_2|\}$ . Since  $v, w_2 \in L_j$ , and  $L_j$  is transitive there is a tile from v to  $w_2$ . We have thus obtained a path in L from  $w_1$  to  $w_2$ , proving that L is chain transitive.

Assume then that L is chain transitive, and consider  $i \neq j$  in  $\{1, 2, \ldots, N\}$ . Let M be the number from Lemma 5.1. There is an  $m \in \mathbb{N}$  such that any path in G of length m or more must pass through at least M consecutive edges in one of the  $G_k$ 's. Let  $w_1 \in L_i, w_j \in L_j$ , both of length 2m. Since L is chain transitive there is a path in L from  $w_1$  to  $w_2$ . Consider the first tile u of the path. This u is the label of path p in G in which at least M consecutive edges among the first m edges must pass through one of the  $G_k$ 's, say  $G_{i_1}$ . Then  $\#(L_i \cap L_{i_1}) = \infty$ , so there is an edge  $i \to i_1$  in  $\mathcal{G}$ , unless  $i = i_1$ . Similarly, among the m last edges in p, at least M consecutive edges must be contained in one of the  $G_k$ 's, say  $G_{i_2}$ . The path p connects therefore the communication classes  $C_{i_1}$  and  $C_{i_2}$ , giving us an edge  $i_1 \to i_2$  in  $\mathcal{G}$ , unless  $i_1 = i_2$ . By repeating these arguments, one tile at the time, we obtain a path  $i \to j$  in  $\mathcal{G}$ , proving that this graph is irreducible.

It is clear from the preceding that there is a finite algorithm which produces the graph of chain connected components of a given labeled graph. We can therefore conclude that it is decidable if the language of a given finite automaton with all states initial and final is the language of finite words occurring infinitely many times in an infinite word.

648

It seems appropriate, as an illustration, to give some simple examples of regular infinite and factorial languages that are chain transitive, and hence can be realized as the language of finite words occurring infinitely many times in an infinite word, and others that are not.



The languages given by the first and third of the labeled graphs can be realized as the language of finite words occurring infinitely many times in an infinite word. The others can not. The chain transitive languages of the first and third of the graphs are not transitive so it follows from Proposition 3.1 that an infinite word x for which the corresponding language becomes equal to  $L_{\infty}(x)$  can not be eventually recurrent.

The reader who is so inclined should have no difficulty in adopting the previous analysis to find necessary and sufficient conditions for a language to be the language of finite words occurring infinitely many times in a *bi-infinite* word.

**Remark 5.3.** In "Open problem 2" of [6], the authors ask if there is any almost periodic infinite word w such that the language of the finite words occurring in w is context-free and not regular. The answer to this question is "No". Assume namely that such a word w exists. Since w is almost periodic, the language of words occurring in w is then also the language of a minimal subshift X, *i.e.* a

#### K. THOMSEN

subshift where all orbits are dense. If the language is context-free, it follows that X must contain a periodic point, *cf.* [1] or [3]. Being minimal, X must then equal the corresponding finite orbit. In particular, the language of X must be regular.

## References

- Y. Bar-Hilel, M. Peres and E. Shamir, On formal properties of simple phrase structure grammar. Z. Phonetik. Sprachwiss. Komm. 14 (1961) 143–172.
- [2] T. Harju and L. Ilie, Languages obtained from infinite words. RAIRO-Inf. Theor. Appl. 31 (1997) 445–455.
- [3] J.E. Hopcroft and J.D. Ullman, Introduction to Automata Theory, Languages and Computation. Addison-Wesley Publishing Company (1979).
- [4] L. Ilie, On subwords of infinite words. Discrete Appl. Math. 63 (1995) 277–279.
- [5] D. Lind and B. Marcus, An Introduction to Symbolic Dynamics and Coding. Cambridge University Press (1995).
- [6] S. Marcus and G. Paun, Infinite (almost periodic) words, formal languages and dynamical systems. Bull. EATCS 54 (1994) 224–231.

Communicated by J. Karhumaki.

Received December 30, 2003. Accepted October, 2004.

To access this journal online: www.edpsciences.org