Languages obtained from infinite words

Tero Harju

University of Turku, Department of Mathematics,
FIN-20014 Turku, Finland

Lucian Ilie

TUCS

Turku Centre for Computer Science
TUCS Technical Report No 19
May 1996

ISBN 951-650-769-7
ISSN 1239-1891

Abstract

Answering two problems formulated by Marcus and Păun, we prove that it is decidable whether or not a context-free language can be written as the set of all finite prefixes of an infinite word and that it is decidable whether or not a regular language can be written as the set of all finite subwords of an infinite word.

TUCS Research Group
Mathematical Structures of Computer Science

1 Introduction and basic definitions

A lot of attention has been devoted recently to the study of infinite words. Several of the main topics considered are: iterative devices generating infinite words [8], [9], the subword complexity of infinite words [1], [6], [7], different ways to relate infinite words and languages of finite words [2], [12], [13].

There are several classical ways to associate a set of finite words to an infinite word α. One can take the set of all finite prefixes or finite subwords of α, $Pref(\alpha)$ or $Sub(\alpha)$, respectively, [12], [13], or the set of all finite words which are not prefixes of α, $Copref(\alpha)$, [2], [3], [4]. Conversely, for a language of finite words L, one can associate infinite words considering the notions of limit [10] or adherence [5].

This paper is devoted to the study of languages obtained from infinite words by taking the set of all finite prefixes or finite subwords, respectively. More precisely, we prove that it is decidable whether or not a context-free language can be written as the set of all finite prefixes of an infinite word and that it is decidable whether or not a regular language can be written as the set of all finite subwords of an infinite word, answering two open problems in [12], [13].

For an alphabet Σ, we denote by Σ^* the set of all finite words over Σ and by Σ^ω the set of all (one-sided) infinite words over Σ; λ denotes the empty word and $\Sigma^+ = \Sigma^* - \{\lambda\}$. For a finite word $w \in \Sigma^*$, we denote by $|w|$ the length of w. For a finite (non-empty) word $w \in \Sigma^+$, we denote by w^ω the infinite word $w^\omega = www\ldots$. For all formal language theory notions and results we refer to [11] and [16].

For an infinite word $\alpha \in \Sigma^\omega$, we denote by $Pref(\alpha)$ the set of all finite prefixes of α. For a language $L \subseteq \Sigma^*$, $Pref(L)$ is the set of all prefixes of words in L. By $Sub(\alpha)$ and $Sub(L)$ we denote the set of all finite subwords of α and, respectively, L. We also denote by $\mathcal{F}_{pref}$ the family of languages of the form $Pref(\alpha)$, for an arbitrary infinite word α, that is,

$$\mathcal{F}_{pref} = \{L \mid \text{ there are } \Sigma \text{ and } \alpha \in \Sigma^\omega \text{ such that } L = Pref(\alpha)\}.$$

By $\mathcal{F}_{sub}$ we denote the corresponding family of languages obtained by taking subwords instead of prefixes.

2 Prefixes of infinite words

In [12] it is proved that it is decidable whether or not a regular language L is in the family $\mathcal{F}_{pref}$ and the case when L is context-free is left open. In what follows, we give a positive answer to this problem.

A language $L \subseteq \Sigma^*$ is called *thin* [14] if and only if, for any $n \geq 0$, L contains at most one string of length n, i.e.,

$$card\{w \in L \mid |w| = n\} \leq 1.$$

We begin with some remarks. (Part of them can be found also in [12] but, for technical reasons, we give the complete proofs.)

Lemma 1 *If L is an infinite language, then L is in the family $\mathcal{F}_{pref}$ if and only if L is thin and L is closed under taking prefixes.*

Proof. If $L \in \mathcal{F}_{pref}$, then, obviously, $L = Pref(L)$ and L is thin (any infinite word has only one prefix of length n, for any $n \geq 0$). Conversely, if $L = Pref(L)$ and L is thin, it follows that, for any $n \geq 0$, L contains exactly one word of the length n, say w_n. If we denote by a_n the last letter of w_n for any $n \geq 1$, then $L = Pref(\alpha)$ for $\alpha = a_1 a_2 a_3 \ldots$. □

Lemma 2 *If $L \subseteq \Sigma^*$ is in the family $\mathcal{F}_{pref}$ then L is context-free if and only if there are two words $u, v \in \Sigma^*$ such that $L = Pref(\alpha)$ for $\alpha = uv^\omega$. Moreover, there is a constant $p = p(L)$ such that $|u| \leq p, |v| \leq p + 1$.*

Proof. Suppose that L is context-free and let $G = (V, \Sigma, S, P)$ be a context-free grammar generating L. Using the Bar-Hillel lemma for G, we get a constant p such that any word $z \in L, |z| > p$, can be written as $z = x_1 x_2 x_3 x_4 x_5$ with $x_2 x_4 \neq \lambda$ and $x_1 x_2^n x_3 x_4^n x_5 \in L$ for any $n \geq 0$. Because $L \in \mathcal{F}_{pref}$, it follows that there is a word $z \in L, |z| = p + 1$. Thus z can be written as $z = x_1 x_2 x_3 x_4 x_5$ with $x_2 x_4 \neq \lambda$ and $x_1 x_2^n x_3 x_4^n x_5 \in L$ for any $n \geq 0$. It is easy to see now that if $x_2 \neq \lambda$, then we can take $u = x_1, v = x_2$ and if $x_2 = \lambda$, then we can take $u = x_1 x_3, v = x_4$ having, in both cases, $L = Pref(uv^\omega)$. We have only to note that, in any case, $|u| \leq p, |v| \leq p + 1$.

The converse is obvious, L being even regular. □

The following result from [15] states the decidability of the thinness problem for context-free languages.

Lemma 3 ([15]) *The thinness problem is decidable for context-free languages.*

We are ready now to prove the main result of this section.

Theorem 4 *It is decidable whether or not an arbitrary context-free language is in the family $\mathcal{F}_{pref}$.*

Proof. Let $L \subseteq \Sigma^*$ be an arbitrary context-free language. Take a context-free grammar $G = (V, \Sigma, S, P)$ generating L and let p be the constant in Bar-Hillel lemma for G. For deciding the above property, we follow the next *algorithm*:

1. Decide whether or not L is thin (this is possible by Lemma 3). If L is not thin, then answer **no** (the answer is correct by Lemma 1).

2. (L is thin) For all pairs of words $(u, v) \in (\Sigma^*)^2$ with $|u| \leq p, |v| \leq p+1$ (so a finite number of choices), decide whether or not $L \subseteq Pref(uv^\omega)$. (That can be decided because $Pref(uv^\omega)$ is a regular language and it is decidable whether or not a context-free language is included in a regular one: $L \subseteq R$ if and only if $L \cap \overline{R} = \emptyset$. For L context-free and R regular, $L \cap \overline{R}$ is effectively context-free and its emptiness is decidable.)

2.1. If a pair (u, v) with the above property was found, then go to step 3.

2.2. If no pair was found, then answer **no** (the answer is correct by Lemma 2).

3. ($L \subseteq Pref(uv^\omega)$ for some $u, v \in \Sigma^*$) We obviously have

$$Pref(uv^\omega) = F \cup \bigcup_{i=0}^{|v|-1} R_i,$$

where

$$F = \{\lambda, u_1, u_1u_2, u_1u_2u_3, \ldots, u_1u_2 \ldots u_{|u|-1}\}$$

is a finite language and

$$\begin{aligned} R_0 &= \{uv^n \mid n \geq 0\}, \\ R_i &= \{uv^n v_1 v_2 \ldots v_i \mid n \geq 0\}, 1 \leq i \leq |v| - 1, \end{aligned}$$

are $|v|$ regular languages.

3.1. Check whether $F \subseteq L$.

3.1.1. In the affirmative case put $L' := u \backslash (L - F)$ and go to step 3.2 (the family of context-free languages is effectively closed under intersection with regular languages ($L - F = L \cap \overline{F}$) as well as under quotient with regular sets).

3.1.2. In the negative case answer **no** (again by Lemma 1, the answer is correct).

3.2. For each $i, 0 \leq i \leq |v| - 1$ (supposing that $v_1 v_2 \ldots v_i = \lambda$ for $i = 0$) do

3.2.1. $L_i := (L' \cap \{v^n v_1 v_2 \ldots v_i \mid n \geq 0\}) / v_1 v_2 \ldots v_i$ (as it was pointed out above, this is effectively constructable).

3.2.2. Define the homomorphism $h : \{a\}^* \longrightarrow \Sigma^*(a \notin \Sigma)$ by $h(a) = v$ and do $L_i' := h^{-1}(L_i)$ (again, the family of context-free languages is effectively closed under inverse homomorphisms).

3.2.3. It follows that $L_i' \subseteq \{a^n \mid n \geq 0\}$ and $R_i \subseteq L$ if and only if $L_i' = \{a^n \mid n \geq 0\}$. But L_i' is a context-free language over a one letter alphabet, hence it is regular and it is decidable whether or not $L_i' = \{a^n \mid n \geq 0\}$.

3.2.4. Decide whether or not $L_i' = \{a^n \mid n \geq 0\}$.

3.2.4.1. If the answer is negative, then answer **no**.

3.2.4.2. If the answer is affirmative, then if $i < |v| - 1$, then $i := i + 1$ and go to step 3.2.

4. Answer **yes**.

If the algorithm has reached step 4, then it follows that for any $i, 0 \leq i \leq |v| - 1$, $R_i \subseteq L$ and, as $F \subseteq L$, we get $L = Pref(uv^\omega)$. Consequently, in all cases, the answer given by the above algorithm is correct. □

3 Subwords of infinite words

In [12] it is proved that it is undecidable whether an arbitrarily given context-free language is in the family $\mathcal{F}_{sub}$ or not. The same problem for regular languages is left open. In the following, we solve this problem in the affirmative.

For a regular language $R \subseteq \Sigma^*$, a finite automaton $\mathcal{A} = (Q, \Sigma, \delta, I, F)$ recognizing R and having a deterministic transition function $\delta : Q \times \Sigma \longrightarrow Q$ is called *strongly minimal* if and only if no state or transition in $\mathcal{A}$ is useless or redundant. That is, if we eliminate any state or transition in $\mathcal{A}$, the obtained automaton recognizes a language strictly contained in R.

Lemma 5 *Any regular language $R \subseteq \Sigma^*$ closed under taking subwords is recognized by a strongly minimal automaton $\mathcal{A} = (Q, \Sigma, \delta, Q, Q)$ in which all states are both initial and final.*

Proof. Let $\mathcal{A}' = (Q', \Sigma, \delta', I, T)$ be a finite deterministic automaton recognizing R. Consider the automaton $\mathcal{A}'' = (Q', \Sigma, \delta', Q', Q')$. Since R is closed under taking subwords, $R = L(\mathcal{A}'')$. Because the equivalence problem is decidable for finite automata, we can now iteratively eliminate from $\mathcal{A}''$ all states and transitions which are either useless or redundant. That is, if $s \in Q'$ (or $\delta'(s, a) = s'$, for $s, s' \in Q, a \in \Sigma$) and the language recognized by the finite automaton $\mathcal{B}$ obtained from $\mathcal{A}''$ by removing the state s together with all transitions containing it (or removing the transition by a from s to s', respectively) is R, then take $\mathcal{B}$ instead of $\mathcal{A}''$ and continue the reduction. Obviously, after a finite number of steps, the automaton $\mathcal{A}$ asked for in the claim of our lemma is obtained. (Note that $\mathcal{A}$ is not necessarily unique, but this will not cause troubles later.) $\square$

For a finite automaton $\mathcal{A} = (Q, \Sigma, \delta, I, F)$, denote by $G(\mathcal{A})$ the graph associated to $\mathcal{A}$ and define the relation $\longrightarrow \subseteq Q \times Q$ by

$$p \longrightarrow q \text{ if and only if there is a path from } p \text{ to } q \text{ in } G(\mathcal{A}).$$

The relation $\equiv \subseteq Q \times Q$ defined by $p \equiv q$ if and only if $p \longrightarrow q$ and $q \longrightarrow p$ is an equivalence relation which induces an acyclic structure on $Q/_\equiv$ i.e., the graph $G = (Q/_\equiv, E)$ with $Q/_\equiv$ as the set of vertices and with the set of edges

$$E = \{([p], [q]) \mid [p], [q] \in Q/_\equiv \text{ and } p \longrightarrow q\}$$

is acyclic. (We have denoted the equivalence class of $p \in Q$ with respect to $\equiv$ by $[p]$.)

The automaton $\mathcal{A}$ is called *diconnected* if and only if there is a state $q \in Q$ such that $Q = [q]$. (For instance, the restriction of any automaton to an equivalence class with respect to the relation $\equiv$ is a diconnected automaton.)

An equivalence class $[p] \in Q/_\equiv$ is called *trivial* if and only if $[p]$ is a singleton ($[p] = \{p\}$) and there is no transition $p \xrightarrow{a} p$, for any $a \in \Sigma$.

A finite automaton $\mathcal{A} = (Q, \Sigma, \delta, I, F)$ is called *ultimately periodic* if and only if there is a state $q \in Q$ such that

$$\begin{array}{rcll} Q - [q] & = & \{s_1, s_2, \ldots, s_k\}, & \text{for some } k \geq 1, \\ [q] & = & \{q_1 = q, q_2, \ldots, q_l\}, & \text{for some } l \geq 1, \end{array}$$

and all transitions in $\mathcal{A}$ are

$$\begin{array}{ll} \delta(s_i, a_i) = s_{i+1}, 1 \leq i \leq k-1, & \text{for some } a_1, a_2, \ldots, a_{k-1} \in \Sigma, \\ \delta(s_k, a_k) = q, & \text{for some } a_k \in \Sigma, \\ \delta(q_j, b_j) = q_{j+1}, 1 \leq j \leq l-1, & \text{for some } b_1, b_2, \ldots, b_{l-1} \in \Sigma, \\ \delta(q_l, b_l) = q, & \text{for some } b_l \in \Sigma. \end{array}$$

Informally speaking, a finite automaton $\mathcal{A}$ is ultimately periodic if and only if $G(\mathcal{A})$ has the form in Figure 1 (using the notations in the definition).

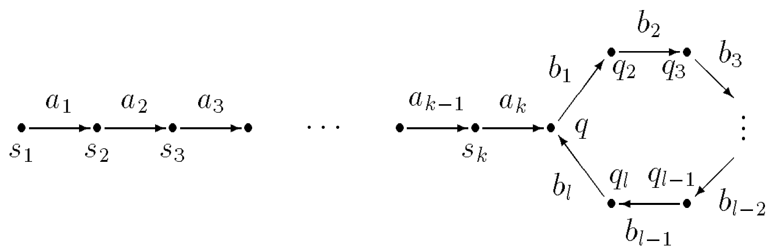


Figure 1

(Note that if $\mathcal{A}$ is ultimately periodic, then $\mathcal{A}$ is not diconnected. Moreover, for any $p \in Q - [q]$, $[p]$ is trivial.)

An equivalence class $[p] \in Q/_{\equiv}$ is called a *source* if and only if there is no $[q] \in Q/_{\equiv} - [p]$ such that $q \longrightarrow p$.

Lemma 6 *For a regular language $R \subseteq \Sigma^*$ such that $R \in \mathcal{F}_{sub}$, let $\mathcal{A} = (Q, \Sigma, \delta, Q, Q)$ be a strongly minimal automaton constructed using Lemma 5 for R. If $\mathcal{A}$ is not diconnected, then there is a unique equivalence class $[p] \in Q/_{\equiv}$ which is a source. Moreover, $[p]$ is trivial and there is exactly one transition leaving p in $\mathcal{A}$.*

Proof. The existance of a source is guaranteed by the fact that $\mathcal{A}$ is not diconnected. (If no source exists, then $\mathcal{A}$ must be diconnected, a contradiction.)

Let us show that any source is trivial. For, take a source $[p] \in Q/_{\equiv}$ and consider the subautomata $\mathcal{A}^{[p]} = ([p], \Sigma, \delta|_{[p]}, [p], [p])$ and $\mathcal{A}^{Q-[p]} = (Q - [p], \Sigma, \delta|_{Q-[p]}, Q - [p], Q - [p])$ of $\mathcal{A}$.

Suppose, contrary to the claim, that $[p]$ is not trivial. Take $p_1, p_2 \in [p]$ such that if $card([p]) \geq 2$, then $p_1 \neq p_2$, othewise $p_1 = p_2 = p$. If $p_1 \neq p_2$, then, as $\mathcal{A}^{[p]}$ is diconnected, there must be in $\mathcal{A}^{[p]}$ a path from p_1 to p_2, say $p_1 \xrightarrow{y} p_2, y \in \Sigma^+$, and another one from p_2 to p_1, say $p_2 \xrightarrow{z} p_1, z \in \Sigma^+$. If $p_1 = p_2$, as $[p]$ is not trivial, there must be a transition $p \xrightarrow{a} p$, for some $a \in \Sigma$, and we can take $y = z = a$. So, in what follows, it will not be important whether $p_1 \neq p_2$ or not. What is important is the fact that, in both cases, the word yz is not empty.

As, obviously, $G(\mathcal{A})$ is connected and the sets of states of $\mathcal{A}^{[p]}$ and $\mathcal{A}^{Q-[p]}$, respectively, are non-empty (p is in $\mathcal{A}^{[p]}$ and, if $\mathcal{A}^{Q-[p]}$ is empty, then $\mathcal{A} = \mathcal{A}^{[p]}$ hence diconnected, a contradiction), there must be a path from a state in $\mathcal{A}^{[p]}$ to one in $\mathcal{A}^{Q-[p]}$ and we can find $p_3 \in [p], q_1 \in Q - [p]$, and a transition $p_3 \xrightarrow{a} q_1, a \in \Sigma$, in $\mathcal{A}$. Since $\mathcal{A}$ is strongly minimal, there is a word $w \in \Sigma^*$ which contains a and is accepted by the automaton $\mathcal{A}$ but not accepted by the automaton obtained from $\mathcal{A}$ by removing the transition $p_3 \xrightarrow{a} q_1$. (That is, when $\mathcal{A}$ accepts w, then it must read a from p_3 to q_1.) It folllows that we can find the states $p_4 \in [p]$, $q_2 \in Q - [p]$ and the words $w_1, w_2 \in \Sigma^*$ such that $w = w_1 a w_2$ and there are paths $p_4 \xrightarrow{w_1} p_3$ in $\mathcal{A}^{[p]}$ and $q_1 \xrightarrow{w_2} q_2$ in $\mathcal{A}^{Q-[p]}$. Using again the fact that $\mathcal{A}^{[p]}$ is diconnected, we get a path from p_1 to p_4 in $\mathcal{A}^{[p]}$, say $p_1 \xrightarrow{u} p_4, u \in \Sigma^*$, see Figure 2 below.

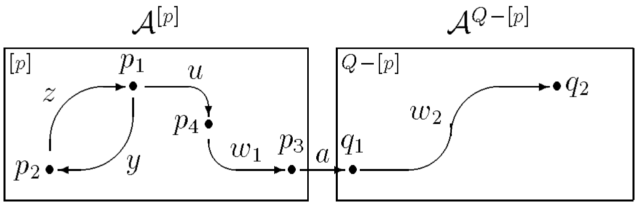
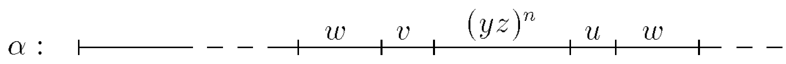


Figure 2

Then $(yz)^*uw_1aw_2 \subseteq R$. Assume $R = Sub(\alpha)$ for some infinite word $\alpha \in \Sigma^\omega$ (there is such an α by hypothesis). As the language $(yz)^*$ contains arbitrarily long words, there must be an $n \geq 0$ such that an occurrence of $(yz)^n uw$ appears in α after an occurrence of w ("after" meaning at a larger distance from the beginning of α). Supposing that we have



for some $v \in \Sigma^*$, α has a subword $wv(yz)^n uw$. But now $wv(yz)^n uw \in R$. Because, when accepting w, $\mathcal{A}$ must read a from p_3 to q_1, using the fact that $[p]$ is a source, we obtain that the word $v(yz)^n uw$ is accepted by $\mathcal{A}$ without reading $p_3 \xrightarrow{a} q_1$. In particular, w is accepted by $\mathcal{A}$ without reading $p_3 \xrightarrow{a} q_1$, a contradiction. It follows that $[p]$ is trivial.

Let us prove now that there is exactly one source in $\mathcal{A}$. For this, suppose that $[p]$ and $[q]$ are two different sources. As shown above, $[p] = \{p\}$ and $[q] = \{q\}$. As $\mathcal{A}$ is strongly minimal, there must be transitions starting from p and q labeled by $a \in \Sigma$ and $b \in \Sigma$, respectively. Moreover, there are some words $ax_1, bx_2 \in L(\mathcal{A})$ such that $\mathcal{A}$ must read the transition which leaves p (q) and is labeled by a (b) in order to accept the word ax_1 (bx_2, respectively). It follows that the word ax_1 cannot be prolonged to the left in R, that is, there is no word $w \in \Sigma^+$ such that $wax_1 \in R$. As $R = Sub(\alpha), \alpha \in \Sigma^\omega$, α must begin with ax_1. Since the same resoning can be done for bx_2, we get that of ax_1 and bx_2 one is a prefix of the other. Suppose that ax_1 is a prefix of bx_2. In this case, ax_1 is accepted by the automaton $\mathcal{A}$ without reading the transition labeled a leaving p, a contradiction.

A similar reasoning shows that for a source $[p]$ the number of transitions leaving p is exactly one. Indeed, it is at least one and if there are two transitions labeled a and b, then $a \neq b$, because $\mathcal{A}$ is deterministic, and then α must start with both a and b, a contradiction. □

Lemma 7 *For an infinite regular language $R \subseteq \Sigma^*$, $R \in \mathcal{F}_{sub}$ if and only if an automaton $\mathcal{A}$ constructed for R in Lemma 5 is either diconnected or ultimately periodic.*

Proof. Suppose first that $\mathcal{A} = (Q, \Sigma, \delta, Q, Q)$ is diconnected and take an arbitrary $q \in Q$. As R is infinite, there must be at least one cycle $q \longrightarrow q$ in $G(\mathcal{A})$ hence there are infinitely many such cycles, the set of them being

$$C(q) = \{w \in \Sigma^* \mid q \xrightarrow{w} q\} = \{w_1, w_2, w_3 \ldots\}.$$

Define

$$\alpha = w_1 w_2 w_3 \ldots.$$

We have $L(\mathcal{A}) = Sub(\alpha)$. The inclusion $Sub(\alpha) \subseteq L(\mathcal{A})$ is trivial. To prove the other one, take $w \in L(\mathcal{A})$. There are $p_1, p_2 \in Q$ such that $p_1 \xrightarrow{w} p_2$. Since $\mathcal{A}$ is diconnected, we have also $q \xrightarrow{u} p_1$, $p_2 \xrightarrow{v} q$, for some $u, v \in \Sigma^*$. It follows that there is an $i \geq 1$ such that $w_i = uwv$, so $w \in Sub(\alpha)$.

If $\mathcal{A}$ is ultimately periodic then $G(\mathcal{A})$ has the form in Figure 1. Hence (using the notations there)

$$L(\mathcal{A}) = Sub(a_1 a_2 \ldots a_k (b_1 b_2 \ldots b_l)^\omega).$$

Coversely, take a regular language $R \subseteq \Sigma^*$ such that $R \in \mathcal{F}_{sub}$ and construct $\mathcal{A}$ as in Lemma 5. If $\mathcal{A}$ is diconnected, then we are done. Otherwise, by Lemma 6, there is exactly one source, say $[p_1]$, in $\mathcal{A}$. Moreover, $[p_1]$ is trivial and there is exactly one transition in $\mathcal{A}$ starting from p_1, say $p_1 \xrightarrow{a_1} p_2$, for some $p_2 \in Q - \{p_1\}$ and $a_1 \in \Sigma$. If $\alpha \in \Sigma^\omega$ with $R = Sub(\alpha)$, then from the proof of Lemma 6, α has the form $\alpha = a_1\alpha'$, $\alpha' \in \Sigma^\omega$.

Denote the automaton obtained from $\mathcal{A}$ by removing p_1 and the transition leaving p_1 (labeled a_1) by $\mathcal{A}_1$ ($\mathcal{A}_1 = (Q - \{p_1\}, \Sigma, \delta|_{Q-\{p_1\}}, Q - \{p_1\}, Q - \{p_1\})$). If $\mathcal{A}_1$ is not diconnected, then the only source in $\mathcal{A}_1$ is $[p_2] = \{p_2\}$. Suppose that the only transition from p_2 is $p_2 \xrightarrow{a_2} p_3, p_3 \in Q - \{p_1, p_2\}, a_2 \in \Sigma$ and put $\mathcal{A}_2 = (Q - \{p_1, p_2\}, \Sigma, \delta|_{Q-\{p_1,p_2\}}, Q - \{p_1, p_2\}, Q - \{p_1 p_2\})$. If $\mathcal{A}_2$ is not diconnected, then we continue our procedure. Obviously, after a finite number of steps, say $k \geq 1$, we get a diconnected $\mathcal{A}_k$. Moreover, $L(\mathcal{A}_k)$ is infinite since R is. It remains to show that $G(\mathcal{A}_k)$ is a cycle. $G(\mathcal{A}_k)$ contains at least one cycle; suppose that there are two distinct cycles (meaning that none of them is contained in the other), the second one being $p_k \xrightarrow{u_k} p_k$. As mentioned, α must start with $a_1 a_2 \ldots a_{k-1}$. We have that $a_1 a_2 \ldots a_{k-1} w_k, a_1 a_2 \ldots a_{k-1} u_k \in Sub(\alpha)$. As $a_1 a_2 \ldots a_{k-1}$ appears only at the beginning of α, it follows that w_k is a prefix of u_k or conversely, a contradiction. $\square$

Because, given a regular language R, a strongly minimal automaton for R is effectively constructable by Lemma 5 and it is decidable whether or not an arbitrary finite automaton is diconnected as well as ultimately periodic, we obtain as a consequence of Lemma 7 the main result of this section.

Theorem 8 *It is decidable whether or not an arbitrary regular language is in the family $\mathcal{F}_{sub}$.*

4 The case of bi-infinite words

A *bi-infinite* (or two-sided) word is an infinite word without any end. (Usually, an one-sided infinite word is viewed as a function $\alpha : \mathbf{N} \longrightarrow \Sigma$, Σ being an alphabet. We can define a bi-infinite word as a function $\alpha : \mathbf{Z} \longrightarrow \Sigma$ or, in fact, as an equivalence class of the set $\Sigma^{\mathbf{Z}}$ with respect to the equivalence relation defined for $\alpha, \beta \in \Sigma^{\mathbf{Z}}$ by $\alpha \sim \beta$ if and only if there is an integer k such that for any $n \in \mathbf{Z}, \alpha(n) = \beta(n+k)$.)

We denote by $\Sigma^{2\omega}$ the set of all bi-infinite words over Σ. For a finite (non-empty) word $w \in \Sigma^+$, we denote by ${}^{\omega}w$ the infinite (to the left) word ${}^{\omega}w = \ldots www$.

Of course one cannot speak about prefixes here but it is still possible to consider finite subwords. We shall denote by $\mathcal{F}_{bi-sub}$ the family of languages of the form $Sub(\alpha)$, for an arbitrary bi-infinite word α, that is,

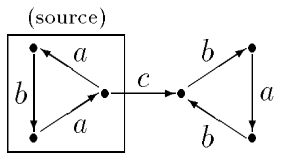
$$\mathcal{F}_{bi-sub} = \{L \mid \text{ there are } \Sigma \text{ and } \alpha \in \Sigma^{2\omega} \text{ with } L = Sub(\alpha)\}.$$

As in the case of one-sided infinite words, it is very easy to prove that it is *undecidable* whether or not an arbitrary context-free language $L \subseteq \Sigma^*$ is in the family $\mathcal{F}_{bi-sub}$ or not (the proof uses the undecidability of the problem of whether $L = \Sigma^*$ or not and is similar to the one of Theorem 6 in [12]).

In what concerns regular languages, we show in this section that the above problem is decidable.

First, note that things are different from the case of one-sided infinite words; for instance, we can find a regular language $R \in \mathcal{F}_{bi-sub}$ such that its automaton $\mathcal{A}$ constructed using Lemma 5 has a non-trivial source (see the picture below)

$\mathcal{A}$:



Obviously, $R = Sub((aba)^* c (bab)^*) = Sub(\alpha)$ for $\alpha = {}^{\omega}(aba)c(bab)^{\omega}$.

Lemma 9 *For a regular language* $R \subseteq \Sigma^*, R \in \mathcal{F}_{bi-sub}$*, let* $\mathcal{A} = (Q, \Sigma, \delta, Q, Q)$ *be the strongly minimal automaton constructed using Lemma 5 for* R*. Then there is a unique equivalence class* $[p] \in Q/_{\equiv}$ *which is a source. Moreover,* $[p]$ *is not trivial.*

Proof. Take $\alpha \in \Sigma^{2\omega}$ such that $R = Sub(\alpha)$.

Suppose that $[p] \in Q/_{\equiv}$ is a source. Then, we have a transition $p \xrightarrow{a} q, a \in \Sigma, q \in Q - [p]$, and a word $w = w'aw'' \in R$ such that $\mathcal{A}$ must read $p \xrightarrow{a} q$ in order to accept w. Since α is bi-infinite, w can be prolonged arbitrarily long to the left in R, hence the language accepted by the subautomaton $\mathcal{A}^{[p]} = ([p], \Sigma, \delta|_{[p]}, [p], [p])$ must be infinite. It follows that $[p]$ is not trivial.

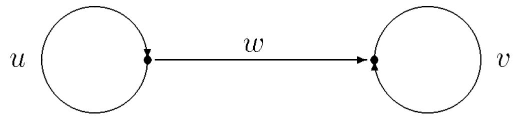
Suppose now that there are two different sources $[p_1], [p_2] \in Q/_{\equiv}$ and the respective transitions and words as above are:

$$p_1 \xrightarrow{a_1} q_1, a_1 \in \Sigma, q_1 \in Q - ([p_1] \cup [p_2]), w_1 = w_1' a_1 w_1'' \in R,$$
$$p_2 \xrightarrow{a_2} q_2, a_2 \in \Sigma, q_2 \in Q - ([p_1] \cup [p_2]), w_2 = w_2' a_2 w_2'' \in R.$$

(So, for $i = 1, 2$, $\mathcal{A}$ must read $p_i \xrightarrow{a_i} q_i$ in order to accept w_i.)

Since w_1 and w_2 appear as subwords of α and $[p_1]$ and $[p_2]$ are sources, the occurrences of $w_1' a_1$ and $a_2 w_2''$ are not overlapped and so are the occurrences of $w_2' a_2$ and $a_1 w_1''$. Consequently, the occurrences of w_1 and w_2 in α are not overlapped and we can find, for instance, $w_1 v w_2 \in Sub(\alpha) = R, v \in \Sigma^*$. But now w_2 can be accepted by $\mathcal{A}$ without reading $p_2 \xrightarrow{a_2} q_2$, a contradiction. The lemma is proved. $\square$

Lemma 10 *For a regular language* $R \subseteq \Sigma^*$*, consider a strongly minimal automaton* $\mathcal{A} = (Q, \Sigma, \delta, Q, Q)$ *accepting* R*. Then* $R \in \mathcal{F}_{bi-sub}$ *if and only if either* $\mathcal{A}$ *is diconnected or there are* $u, v, w \in \Sigma^*$ *such that* $G(\mathcal{A})$ *has the form*



Proof. If $\mathcal{A}$ is diconnected, then we can prove as in Lemma 7 that $R \in \mathcal{F}_{bi-sub}$.

If $G(\mathcal{A})$ has the form in the statement, then $R = Sub(\alpha)$ for $\alpha = {}^{\omega}uwv^{\omega} \in \Sigma^{2\omega}$.

Conversely, suppose that $R \in \mathcal{F}_{bi-sub}$. If $\mathcal{A}$ is not diconnected, then, by Lemma 9, we get a $p \in Q$ such that $[p]$ is the only source in $\mathcal{A}$ and $[p]$ is not trivial. If $\mathcal{A}^{[p]} = ([p], \Sigma, \delta|_{[p]}, [p], [p])$, then, as in the proof of Lemma 7, one can show that

(i) $G(\mathcal{A}^{[p]})$ is a cycle (u in the figure above),

(ii) $G(\mathcal{A}^{Q-[p]})$ is either of the form in Figure 1 or a cycle.

In both cases, the form of $\mathcal{A}$ in the statement of our lemma is obtained.
$\square$

The main theorem of this section is a consequence of Lemmas 9 and 10.

Theorem 11 *It is decidable whether or not an arbitrary regular language is in the family* $\mathcal{F}_{bi-sub}$.

References

[1] J.-P. Allouche, Sur la complexité des suites infinies, *Bull. Belg. Math. Soc.*, 1(1994), 133–143.

[2] J.-M. Autebert, J. Gabarro, Iterated GSM's and Co-CFL, *Acta Informatica*, 26(1989), 749–769.

[3] J.-M. Autebert, P. Flajolet, J. Gabarro, Prefixes of infinite words and ambiguous context-free languages, *Inf. Process. Lett.*, 25(1987), 211-216.

[4] J. Berstel, Every iterated morphism yields a Co-CFL, *Inf. Process. Lett.*, 22(1986), 7–9.

[5] L. Boasson, M. Nivat, Adherence of languages, *J. Comput. Syst. Sci.*, 20(1980), 285–309.

[6] J. Cassaigne, *Motifs évitables et régularités dans les mots*, Thèse de Doctorat, Université Paris 6, LITP research report TH 94-04.

[7] J. Cassaigne, J. Karhumäki, Toeplitz words, generalized periodicity and periodically iterated morphisms, submitted, 1995.

[8] K. Culik II, J. Karhumäki, Iterative devices generating infinite words, *IJFCS*, 1(1994), 69–97.

[9] K. Culik II, J. Karhumäki, A. Lepistö, Alternating iteration of morphisms and the Kolakoski sequence, in: G. Rozenberg, A. Salomaa (eds.), *A memorial volume for A. Lindenmayer* (Springer-Verlag, 1992), 93–103.

[10] S. Eilenberg, *Automata, languages and machines*, Vol 1, Academic Press, New York, 1974.

[11] J. E. Hopcroft, J. D. Ullman, *Introduction to automata theory, languages, and computation*, Addison-Wesley, Reading, Mass., 1979.

[12] S. Marcus, Gh. Păun, Infinite (almost periodic) words, formal languages and dynamical systems, *The Bulletin of the EATCS*, 54(1994), 224–231.

[13] S. Marcus, Gh. Păun, Infinite words and their associated formal languages, in: C. Calude, M. J. J. Lennon, H. Maurer (eds.), *Salodays in Auckland* (Auckland Univ. Press, 1994), 95–99.

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

[15] D. Raz, Length consideration in formal language theory, *Theoretical Computer Science*, to appear.

[16] A. Salomaa, *Formal Languages*, Academic Press, New York, 1973.

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.abo.fi

University of Turku
- **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 Science**