Factors of words [1]

Danièle Beauquier and Jean-Eric Pin,

LITP, Université Paris VI et CNRS.

The first motivation of this paper was to solve the following open problem proposed by Parikh to the first author : if two biinfinite words have the same set of factors, do they satisfy the same *first* order formulas in the theory of successor ?[2] Since it is well known that the recognizable sets of biinfinite words can be defined in the monadic *second* order theory of successor, the problem of Parikh leads naturally to the following question : what kind of recognizable sets of biinfinite words can be defined by a first order formula in the theory of successor ? We solve both questions in this paper, but the search of these problems lead us to the solution of several other interesting problems on factors of words.

We first consider factors of *finite* words, which have been used for a long time in language theory. Everyone knows the *local languages* which occur for instance in the theorem of Chomsky-Schützenberger on context-free languages. Roughly speaking, the words of a local language are described by their first and last letter, and by their factors of length 2. For instance, the language $(ab)^+$ is the set of all words whose first letter is "a", whose last letter is "b", and containing no occurrence of aa and bb. The *locally testable languages* generalize local languages : the membership of a given word in such a language is determined by the presence or absence in this word of factors of a fixed length k (the order in which these factors occur and their frequency is not relevant), and by the prefixes and suffixes of length < k of the word. In

---

[1] This work was supported by the PRC Mathématique et Informatique.

[2] In order to avoid any confusion, we mention that the corresponding problem for the logic containing a symbol "<" has been solved by Perrin and Schupp. But the problem of Parikh is different.

terms of automata, this corresponds to finite automata equipped with a "sliding" window of size k through which the word is scanned. The locally testable languages are characterized by a deep and nice algebraic property of their syntactic semigroup, discovered independently by Brzozowski-Simon [6] and McNaughton [15].

There are several possible variations around this definition. First, one can drop the conditions about the prefixes and suffixes. Membership in this type of languages, that we call *strongly locally testable*, is determined only by factors of a fixed length k. For instance, if A = {a,b,c,d}, the language $c(ab)^+d$ is the set of all words whose set of factors of length 2 is either {ca,ab,bd} or {ca,ab,ba,bd}. Surprinsingly, this rather natural family of rational languages does not seem to have been considered previously in the literature. We show that this family is also decidable and characterized by another nice algebraic property of the syntactic semigroup.

A second natural extension is to take in account the number of occurrences of the factors of the word. However, since we want to use finite automata to recognize our languages, we shall just count factors up to a certain threshold. Threshold counting is the favorite way of counting of small children : they can distinguish 0, 1, 2, ... but after a certain number n (the threshold), all numbers are "big". From a more mathematical point of view, two positive integers s and t are congruent threshold n if $s = t$ or if $s \geq n$ and $t \geq n$ (another possibility, not considered in this paper, would be to count modulo some fixed integer m). We call the languages obtained in this way *locally threshold testable* (LTT) - respectively *strictly locally threshold testable* (SLTT) if we ignore the conditions on the prefixes and suffixes. A deep result of Thérien and Weiss [26] yields a syntactic characterization of LTT languages. In view of these results, it is reasonable to think that the family of strongly locally threshold testable sets is also decidable. However, we have not yet a proof of this fact. On the other hand, Thomas [30] proved that a language is LTT if and only if it can be defined in the first order theory of successor.

One can also use factors to define sets of infinite words. For instance, Pécuchet [18] has defined locally testable sets of infinite words. Extending this definition, we define the families of (strongly) locally (threshold) testable sets of biinfinite words. Of course, since there is no hope to define a prefix or a suffix of a biinfinite word, the definition has to be adapted. Instead of prefixes (suffixes) one considers the words occurring "infinitely many times on the left (right)". We show that, with this definition, one recovers all the algebraic characterizations with syntactic semigroups that were true for finite words. That is, given a rational (or regular) set of biinfinite words, one can effectively decide whether it is locally testable, strongly locally testable or locally threshold testable.

An interesting particular case is the set S(u) of all the shifts of a given biinfinite word u. It is a well-known fact that this set is recognizable if and only if u is of the form $^{\omega}fgh^{\omega}$. In fact, in this case, a much stronger result is true : S(u) is SLTT, and it suffices to count the factors threshold 2 to determine membership in S(u). If furthermore, f and h are non conjugate primitive words, or if f = g = h, then S(u) is strongly locally testable.

The connections with logic mentionned above can also be extended. Indeed, a set of biinfinite words is definable by a first order formula if and only if it is strongly locally threshold testable. This follows essentially from a back and forth argument of game theory due to Thomas [27]. This result, together with the description of S(u) given above, leads to a positive answer to the problem of Parikh.

Due to the lack of place, the proofs have been omitted and will be published elsewhere. However, we would like to mention that our results concerning sets of finite words are already non-trivial and their extension to biinfinite words requires a number of technical subtleties. Although these results may accredit the idea that "well, everything works for (bi)infinite words just like for finite words", this is in fact not the rule, and Pécuchet has given a number of counterexamples to this too optimistic belief.

1. Preliminaries.

1.1 Words.

Let A be a finite alphabet. We denote by A^* the set of words over A, and by A^+ the set of non empty words. If u is a word of length $\geq k$, we denote by $p_k(u)$ and $s_k(u)$, respectively, the prefix and suffix of length k of u. A factor x of a word u is *strict* if there exist $s, t \in A^+$ such that u = sxt. If u and x are two words, we denote by $\begin{bmatrix} u \\ x \end{bmatrix}$ the number of occurrences of the factor x in u.

A biinfinite word is an element of $A^{\mathbb{Z}}$, that is, an application from $\mathbb{Z}$ to A. A factor x of a biinfinite word u occurs "infinitely many on the right (respectively left) of $u = ...a_{-2}a_{-1}a_0a_1a_2$ ... " if for every $n > 0$ (respectively $n < 0$), there exists $m \geq n$ (respectively $m+k \leq n$) such that is $a_m ... a_{m+k} = x$. A biinfinite word u is *recurrent* if every factor of u occurs infinitely many on the right and on the left. Given a language L of A^*, we denote by $\overleftrightarrow{L}$ the set of biinfinite words $u = ...a_{-2}a_{-1}a_0a_1a_2 ...$ such that there exist two increasing sequences of integers $(m_k)_{k>0}$ and $(n_k)_{k>0}$ such that $a_{-m_k} ... a_{n_k} \in L$ for every $k > 0$.

1.2 Finite semigroups.

An element e of a semigroup S is *idempotent* if $e^2 = e$. The set of idempotents of a semigroup S is denoted by E(S). A non-empty finite semigroup S always contains at least one idempotent. Recall the definitions of the Green's relations $\mathcal{R}$, $\mathcal{L}$ and $\mathcal{D}$:

s $\mathcal{R}$ t if and only if there exists $u, v \in S^1$ such that su = t and tv =s,

s $\mathcal{L}$ t if and only if there exists $u, v \in S^1$ such that us = t and vt =s,

s $\mathcal{D}$ t if and only if there exists $u \in S^1$ such that s $\mathcal{R}$ u and u $\mathcal{L}$ t.

The next lemma is the cornerstone of most results on infinite words.

Lemma 1.1. Let A be an alphabet (finite or infinite), and let $\varphi : A^+ \to S$ be a semigroup morphism into a finite semigroup. Then, for every $u \in A^{\mathbb{Z}}$, there exists $(e, s, f) \in S^3$ such that $e^2 = e$, $f^2 = f$, es = s = sf and $u \in {}^{\omega}(e\varphi^{-1})(s\varphi^{-1})(f\varphi^{-1})^{\omega}$.

This lemma leads to the following definition : a triple (e, s, f) $\in S^3$ such that $e^2 = e$, $f^2 = f$, $es = s = sf$ will be called a *linked triple* .

1.3. Recognizable sets of $A^{\mathbb{Z}}$.

Let S be a finite semigroup and let $\varphi : A^+ \to S$ be a semigroup morphism. A subset L of $A^{\mathbb{Z}}$ is *φ-simple* if it is of the form

$$\omega(e\varphi^{-1})(s\varphi^{-1})(f\varphi^{-1})^\omega$$

where (e, s, f) is a linked triple.

Definition. A subset L of $A^{\mathbb{Z}}$ is *recognized* by φ if and only if L is finite union of φ-simple sets. It is *saturated* by φ if and only if, for every φ-simple set X, either $X \subset L$ or $X \subset A^{\mathbb{Z}} \setminus L$. This is equivalent to say that if $X \cap L \neq \emptyset$, then $X \subset L$.

If φ saturates L, then it recognizes L, but the converse is not true. Furthermore, if φ recognizes a language L of A^+ (in the usual sense), then it also recognizes $\overleftrightarrow{L}$.

If L is recognized by a semigroup morphism $\varphi : A^+ \to S$, we call *the image of L by φ* the set of linked triples (e,s,f) such that $\omega(e\varphi^{-1})(s\varphi^{-1})(f\varphi^{-1})^\omega \cap L \neq \emptyset$.

One can show [22, 19] that if L is a recognizable subset of $A^{\mathbb{Z}}$, there exists a smallest finite semigroup which saturates L. This semigroup is called the *syntactic semigroup* of L and is denoted by S(L). The morphism $\eta : A^+ \to S(L)$ which saturates L is the *syntactic morphism* of L. Finally, the image P(L) of L is the *syntactic image* of L. Note that

$$L = \bigcup_{(e,\, s,\, f) \in P(L)} \omega(e\eta^{-1})(s\eta^{-1})(f\eta^{-1})^\omega.$$

Let **V** be a variety of finite semigroups and let $\mathcal{V}$ be the variety of languages corresponding to **V**. The following proposition extends to biinfinite words the corresponding result of [18] for infinite words.

Proposition 1.2. Let L be a subset of $A^{\mathbb{Z}}$. Then the following conditions are equivalent :

(1) L is recognized by a semigroup of **V**,

(2) L is finite union of sets of the form $^{\omega}XYZ^{\omega}$, where X^+, Z^+, X^*YZ^* are in $A^+\mathcal{V}$,

(3) L is recognized by a finite automaton whose transition semigroup belongs to **V**.

2. Factors of finite words.

Threshold counting can be used to define some important families of recognizable languages. Let $x \in A^+$ and let $k \geq 0$. We set

$$L(x, k) = \{ u \in A^+ \mid u \text{ contains at least k occurrences of } x \}.$$

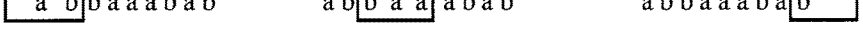
For every n, $t > 0$, one can now define a congruence $\sim_{n,t}$ of finite index on A^+ by setting $u \sim_{n,t} v$ if and only if

(a) u and v have the same prefixes of length < n,

(b) u and v have the same suffixes of length < n,

(c) for every word x of length ≤ n, and for every $k \leq t$, u belongs to L(x, k) if and only if v belongs to L(x, k).

Intuitively, two words u and v are congruent modulo $\sim_{n,t}$ if they cannot be distinguished by a finite automaton equipped with threshold t counters (to count the number of occurrences of each factor of length n) and a window of size n to scan the word. The window can also be moved beyond the first and last letter of the word, so that the prefixes and suffixes of length < n can be read. For instance, if n = 3, and u = abbaaabab, different positions of the window are represented on the following diagrams :



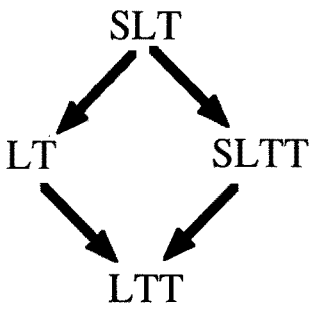
When t = 1, we shall use the notation $\sim_n$ instead of $\sim_{n,1}$. Thus two words are $\sim_n$-equivalent if they have the same prefixes and suffixes of length < n and the same factors of length n (without counting multiplicity). According to [6], we say that a language L of A^+ is *locally n-testable* if L is union of $\sim_n$-classes. It is equivalent to say that L is recognized by the semigroup morphism $\pi_n : A^+ \to S_n = A^+/\sim_n$.

A language is *locally testable* if it is locally n-testable for some $n > 0$. Equivalently, a language is locally testable if and only if it is a boolean combination of languages of the form uA^*, A^*v or A^*wA^* where $u, v, w \in A^+$. For instance, if $A = \{a,b\}$, the language $(ab)^+$ is locally testable since

$$(ab)^+ = (aA^* \cap A^*b) \setminus (A^*aaA^* \cup A^*bbA^*).$$

More generally, we say that a language L of A^+ is *locally threshold t, n-testable* if L is union of $\sim_{n,t}$-classes. It is *locally threshold t testable* if it is locally threshold t n-testable for some $n > 0$. Finally, L is *locally threshold testable* (LTT) if it is locally threshold t testable for some $t > 0$. Equivalently, a language is locally threshold testable if and only if it is a boolean combination of languages of the form uA^*, A^*v or $L(x,k)$ where $u, v, x \in A^+$ and $k > 0$.

If we do not allow the window to move beyond the limits of the word, our machine can count the number of occurrences of a each factor of length $\leq n$, but cannot specify the prefixes (respectively suffixes) of length $< n$. This leads to the following definitions. A language L of A^+ is *strongly locally testable* (SLT) if it is a boolean combination of languages of the form A^*wA^* where $w \in A^+$. A language L of A^+ is *strongly locally threshold testable* (STLT) if it is a boolean combination of languages of the form $L(x,k)$ where $x \in A^+$ and $k > 0$. Note that SLT and STLT languages can also be defined in terms of equivalences (it suffices to remove conditions (a) and (b) in the definition of $\sim_{n,t}$) but these equivalences are no longer congruences. The various inclusions between these families are summarized in the following diagram.



All these inclusions are proper : for instance, $(ab)^+$ is LT (and LTT) but is not SLT (nor SLTT), a^*ba^* is LTT (and SLTT) but is not LT (nor SLT).

We now give effective characterizations for three of these families of languages. In order to keep a standard notation for the subsequent statements, we shall denote by L a recognizable language of A^+, by S(L) the syntactic semigroup of L, by $\eta : A^+ \to S(L)$ the syntactic morphism of L, and by $P(L) = L\eta$ the syntactic image of L. Recall that a finite semigroup S is *locally idempotent and commutative* if, for every idempotent e of S, the subsemigroup eSe is idempotent and commutative. Equivalently, S is locally idempotent and commutative if, for every e, s, $t \in S(L)$ such that $e = e^2$,

$$(ese)^2 = (ese) \text{ and } (ese)(ete) = (ete)(ese).$$

We can now state

Theorem 2.1. [6, 15] Let L be a recognizable language of A^+, and let n be the cardinal of S(L). Then the following conditions are equivalent :

(1) L is locally testable,

(2) L is locally n-testable,

(3) S(L) is locally idempotent and commutative.

SLT-languages do not form a variety of languages in the sense of Eilenberg. However, they also have a syntactic characterization, but a condition on P(L) is required. Let S be a finite semigroup and let P be a subset of S. We say that P saturates the $\mathcal{D}$-classes of S if, for every $\mathcal{D}$-class D of S, $D \cap P \neq \emptyset$ implies that D is contained in P. It is equivalent to say that $s \in P$ and $s \mathcal{D} t$ imply $t \in P$. We can now state our syntactic characterization of SLT-languages.

Theorem 2.2. A language L is strongly locally testable if and only if S(L) is locally idempotent and commutative and P(L) saturates the $\mathcal{D}$-classes of S(L).

The syntactic characterization of locally threshold testable languages requires some new definitions. Given a semigroup S, we form a graph G(S) as follows. The vertices of G(S) are

the idempotents of S, and the edges from e to f are the elements of the form esf. We denote by **V** the variety of all aperiodic semigroups S the graph of which satisfies the condition :

if p and r are edges from e to f, and if q is an edge from f to e, then pqr = rqp.

A slight modification of the arguments of [26] leads to the following statement.

Theorem 2.3. A language L is locally threshold testable if and only if S(L) belongs to the variety **V**.

Sketch of the proof. It is not too difficult to show, for instance by imitating the proof of theorem 2.1 given in Eilenberg's book, that a language is LTT if and only if its syntactic semigroup belongs to the variety **ACom** * **LI**, where **ACom** denotes the variety of commutative and aperiodic monoid, **LI** denotes the variety of locally trivial semigroup, and * is the semidirect product of varieties. Now, this condition is not effective, but the paper of Thérien and Weiss [26] is precisely devoted to the study of the varieties of the form **W** * **LI** and the arguments of this paper show in particular that **ACom** * **LI** = **V**. □

The next statement summarizes the results of this section.

Corollary 2.4. For a given recognizable subset L of A^+, the following properties are effectively decidable :

(1) L is locally testable,

(2) L is strongly locally testable,

(3) L is locally threshold testable,

We conjecture that one can also decide whether L is SLTT. It is natural to guess, in view of the previous results, that L is SLTT if and only if S(L) $\in$ **V** and P(L) saturates the $\mathcal{D}$-classes of S(L). Unfortunately, these conditions are necessary, but are not sufficient : the language $\{1,c\}(ac)^+(bc)^+(ac)^*a\{1,c\}$ satisfies these conditions, but is not SLTT.

3. Threshold counting on biinfinite words.

We first extend the definition of section 2 to biinfinite words. Let $x \in A^+$ and let $k \geq 0$. We set

$F(x, k) = \{ u \in A^{\mathbb{Z}} \mid$ u contains at least k occurrences of x $\}$.

$R(x, \infty) = \{ u \in A^{\mathbb{Z}} \mid$ u contains infinitely many occurrences of x on the right $\}$.

$L(x, \infty) = \{ u \in A^{\mathbb{Z}} \mid$ u contains infinitely many occurrences of x on the left $\}$.

Note that $F(x, 1) = {}^{\omega}AxA^{\omega}$. A subset L of $A^{\mathbb{Z}}$ is *locally n-testable* if it is a boolean combination of languages of the form ${}^{\omega}AxA^{\omega}$, $R(x, \infty)$ or $L(x, \infty)$, where $|x| \leq n$. It is *locally testable* if it is *locally n-testable* for some $n > 0$. It is *strongly locally testable* (SLT) if it is a boolean combination of languages of the form ${}^{\omega}AxA^{\omega}$ where $x \in A^+$.

More generally, a subset L of $A^{\mathbb{Z}}$ is *locally threshold t testable* if it is a boolean combination of sets of the form $F(x, k)$, $R(x, \infty)$ or $L(x, \infty)$, where $x \in A^+$ and $k \leq t$. It is *locally threshold testable* (LTT) if it *locally threshold t testable* for some $t > 0$. It is *strongly locally threshold t testable* if it is a boolean combination of languages of the form $F(x, k)$ where $x \in A^+$ and $k \leq t$. It is *strongly locally threshold testable* (SLTT) if it is *strongly locally threshold t testable* for some $t > 0$.

We first give an effective description of locally testable sets of $A^{\mathbb{Z}}$. A similar result has been obtained by Pécuchet [18] for infinite words and the proof for $A^{\mathbb{Z}}$ follows essentially the same arguments (with some technical complications).

Theorem 3.1. Let L be a recognizable subset of $A^{\mathbb{Z}}$. Then the following conditions are equivalent :

(1) L is locally testable,

(2) L is locally 2n-testable, where $n = |S(L)|$,

(3) L is recognized by a locally idempotent and commutative semigroup,

(4) L is saturated by a locally idempotent and commutative semigroup,

(5) S(L) is locally idempotent and commutative,

(6) L is a boolean combination of bilimits of locally testable languages.

Let S be a finite semigroup and let P be a subset of the set

$$T_S = \{ (e, s, f) \mid e = e^2, f = f^2 \text{ and } es = s = sf \}.$$

We say that P saturates the $\mathcal{D}$-classes of S if for every (e, s, f), (e', s', f') $\in T_S$, the conditions (e, s, f) $\in$ P and s $\mathcal{D}$ s' imply (e', s', f') $\in$ P.

If L is a recognizable subset of $A^{\mathbb{Z}}$, we denote by $\eta : A^+ \to S(L)$ its syntactic morphism and by P(L) the image of L by η.

Theorem 3.2. Let L be a recognizable subset of $A^{\mathbb{Z}}$. Then the following conditions are equivalent :

(1) L is strongly locally testable,

(2) L is a boolean combination of bilimits of strongly locally testable languages,

(3) S(L) is locally idempotent and commutative and P(L) saturates the $\mathcal{D}$-classes of S(L).

Theorem 3.3. Let L be a recognizable subset of $A^{\mathbb{Z}}$. Then the following conditions are equivalent :

(1) L is locally threshold testable,

(2) L is recognized by a semigroup of **V**,

(3) L is saturated by a semigroup of **V**,

(4) S(L) belongs to **V**,

(5) L is a boolean combination of languages of the form $^{\omega}XYZ^{\omega}$, where X^+, Z^+, X^*Y and YZ^* are locally threshold testable languages.

(6) L is a boolean combination of bilimits of locally threshold testable languages.

Corollary 3.4. For a given recognizable subset L of $A^{\mathbb{Z}}$, the following properties are effectively decidable :

(1) L is locally testable,

(2) L is strongly locally testable,

(3) L is threshold locally testable,

The set of shifts S(u) of a given word u is not always recognizable. More precisely, one has the following theorem.

Theorem 3.5. Let $u \in A^{\mathbb{Z}}$. Then the following conditions are equivalent :

(1) S(u) is recognizable,

(2) S(u) is locally threshold testable,

(3) S(u) is strongly locally threshold 2 testable,

(4) $S(u) = {}^{\omega}fgh^{\omega}$ for some $f, g, h \in A^+$.

Thus, if S(u) is recognizable, it can be entirely described by counting the factors of u in the mode "zero, one, many". Of course, there are analoguous results for locally testable and strongly locally testable sets.

Theorem 3.6. Let $u \in A^{\mathbb{Z}}$. Then the following conditions are equivalent :

(1) S(u) is locally testable,

(2) S(u) is strongly locally testable,

(3) $S(u) = {}^{\omega}g^{\omega}$ for some $g \in A^+$ or $S(u) = {}^{\omega}fgh^{\omega}$ where $g \in A^*$ and f and h are non conjugate primitive words of A^+.

Proof. The equivalence of (2) and (3) is shown in [5]. Since (2) implies (1), it remains to show that (1) implies (3). After the results above, it suffices to show that if $S(u) = {}^{\omega}fgf^{\omega}$ for some $f, g \in A^+$ such that f is primitive and g is not a power of f, then S(u) is not locally testable. But this is clear, since for every $n > 0$, any word u of ${}^{\omega}fgf^{\omega}$ and any word v of ${}^{\omega}fgf^{n}gf^{\omega}$ have the same left recurrent (respectively right recurrent) factors and the same factors of length n. □

For instance, the set ${}^{\omega}aba^{\omega}$ is strongly locally threshold 2 testable (it is the set of all words containing exactly one occurrence of "b") but it is not locally testable.

4. Expressive power of first order logic.

We first review the results corresponding to the first order theory of linear ordering, due to Ladner, Thomas and Perrin [11, 20, 23, 28] .

Theorem 4.1. Let L be a subset of $A^{\mathbb{Z}}$. Then the following conditions are equivalent :

(1) L is definable in $Th_1(<, (R_a)_{a \in A})$,

(2) L is a star-free set of $A^{\mathbb{Z}}$,

(3) L is a boolean combination of bilimits of star-free languages,

(4) the syntactic semigroup of L is aperiodic.

The following theorem gives the corresponding results for the first order theory of successor.

Theorem 4.2. Let L be a subset of $A^{\mathbb{Z}}$. Then the following conditions are equivalent :

(1) L is strongly locally threshold testable,

(2) L is definable in $Th_1(S, (R_a)_{a \in A})$,

(3) L is definable by a formula of $Th_1(S, (R_a)_{a \in A})$ containing only existential quantifiers.

Proof. Clearly (3) implies (2). We show that (1) implies (3). Let $x = a_1 \dots a_m \in A^+$. Then for every $k > 0$, the set F(x, k) is defined by the existential formula

$$\psi = \exists n_1 \exists n_2 \dots \exists n_k \left(\bigwedge_{1 \le i,j \le k} \neg (n_i = n_j) \right) \wedge \varphi(n_1) \wedge \varphi(n_2) \wedge \dots \wedge \varphi(n_k)$$

$$\text{where } \varphi(n) = \left(R_{a_1} n \wedge R_{a_2}(n+1) \wedge \cdots \wedge R_{a_k}(n+m-1) \right).$$

If L is strongly locally threshold testable, L is a boolean combination of languages of the form F(x, k) and thus L can be expressed by a formula containing only existential quantifiers.

The more difficult implication, (2) implies (1), follows from an argument of game theory, given in [27].

5. Separative power of first order logic.

The following result describes the separative power of the first order theory of linear ordering and is mainly a consequence of the work of Perrin and Schupp [24]:

Theorem 5.1. Let $u, v \in A^{\mathbb{Z}}$. Then the following conditions are equivalent :

(1) either u and v are shift-equivalent, or u and v are recurrent with the same set of factors,

(2) u and v are equivalent in $Th_1(<, (R_a)_{a \in A})$,

(3) u and v are equivalent in $Th_2(S, (R_a)_{a \in A})$.

(4) u and v are equivalent in $Th_2(<, (R_a)_{a \in A})$.

For the theory of successor, the corresponding result is rather different.

Theorem 5.2. Let $u, v \in A^{\mathbb{Z}}$. Then the following conditions are equivalent :

(1) u and v have the same sets of factors,

(2) u and v are equivalent in $Th_1(S, (R_a)_{a \in A})$,

(3) u and v satisfy the same existential formulas of $Th_1(S, (R_a)_{a \in A})$.

This solves positively the problem proposed by Parikh.

Proof. Clearly (2) implies (3). We show that (3) implies (1). Let $x = a_1 \dots a_k$ be a word. Then x is a factor of u (respectively v) if and only if u (v) satisfies the existential formula

$$\varphi_x = \exists\, n \ \left(R_{a_1} n \wedge R_{a_2}(n+1) \wedge \cdots \wedge R_{a_k}(n+k-1)\right)$$

Since u and v satisfy the same existential formulas, x is a factor of u if and only if it is a factor of v.

Finally, we show that (1) implies (2). Let φ be a sentence of $Th_1(S, (R_a)_{a \in A})$ and let L be the set of biinfinite words satisfying φ. By Theorem 4.2, L is STLT, that is, is a boolean combination of sets of the form F(x, k) where $x \in A^+$ and $k > 0$. Therefore, it suffices to show that $u \in F(x, k)$ is equivalent to $v \in F(x, k)$. Assume that $u \in F(x, k)$. Then u contains a factor w containing at least k occurrences of x. Now by (1), w is a factor of v, so that $v \in F(x, k)$. A

dual argument would show that $v \in F(x, k)$ implies $u \in F(x, k)$, and this concludes the proof. □

References.

[1] A. Arnold, A syntactic congruence for rational ω-languages, *Theor. Comput. Sc.* **39**, 1985, 333-335.

[2] D. Beauquier, Bilimites de langages reconnaissables, *Theor. Comput. Sc.* **33**, 1984, 335-342.

[3] D. Beauquier, Ensembles reconnaissables de mots biinfinis, limite et déterminisme *in Automata on Infinite Words* (M. Nivat and D. Perrin, Eds.), Lecture Notes in Computer Science **192**, 1985, 28-46.

[4] D. Beauquier et M. Nivat, About rational sets of factors of a biinfinite word. *Proc. 12th ICALP* (W.Brauer, Ed.), Lecture Notes in Computer Science **194**, 1985, 33-42.

[5] D. Beauquier, Thin homogeneous sets of factors, Foundations of Software Technology and Theoretical Computer Science, Lecture Notes in Computer Science **241**, 239-251.

[6] J.A. Brzozowski and I. Simon, Characterizations of locally testable languages, *Discrete Math.* **4**, 1973, 243-271.

[7] J.R. Büchi, On a decision method in restricted second order arithmetic, *in Logic Methodology and Philosophy of Science*, (Proc. 1960 Internat. Congr.), Standford University Press, stanford, California, 1960, 1-11.

[8] K.J. Compton, On rich words, *in Combinatorics on words, Progress and Perspectives* (L.J. Cummings ed.) Academic Press 1983, 39-61.

[9] H.D. Ebbinghaus, J. Flum, W. Thomas, *Mathematical Logic,* Springer, 1984.

[10] S. Eilenberg, *Automata, Languages and Machines,* Academic Press, New York, Vol. A, 1974; Vol B, 1976.

[11] R. Ladner, Application of model-theoretic game to discrete linear orders and fininte automata, *Information and Control* **33**, 1977, 281-303.

[12] G. Lallement, *Semigroups and Combinatorial Applications,* Wiley, New York, 1979.

[13] M. Lothaire, *Combinatorics on words,* Encyclopedia of Mathematics 17, Addison-Wesley, New York, 1983.

[14] R. McNaughton, Testing and generating infinite sequences by a finite automaton, *Information and Control* **9**, 1971, 521-530.

[15] R. McNaughton, Algebraic decision procedures for local testability, *Math. Syst. Theor.* **8**, 1974, 60-76.

[16] M. Nivat and D. Perrin, Ensembles reconnaissables de mots biinfinis, *Canad. J. Math.* **38**, 1986, 513-537.

[17] J.P. Pécuchet, On the complementation of Büchi automata, *Theor. Comput. Sci.* **47**, 1986, 95-98.

[18] J.P. Pécuchet, Variétés de semigroupes et mots infinis, *Proc. 3rd Symp. on Theor. Aspects of Comp. Sci.* (G. Vidal-Naquet, Ed.), Lecture Notes in Computer Science **210**, 1986, 180-191.

[19] J.P. Pécuchet, Etude syntactique des parties reconnaissables de mots infinis, *Proc. 13th ICALP* (L. Kott, Ed.), Lecture Notes in Computer Science **226**, 1986, 294-303.

[20] D. Perrin, Recent results on automata and infinite words, *Math. Found. of Comp. Sci.* (M.P. Chytil, V. Koubek, Eds.), Lecture Notes in Computer Science **176**, 1984, 134-148.

[21] D. Perrin, Variétés de semigroupes et mots infinis, *C. R. Acad. Sci. Paris* **295**, 1985, 595-598.

[22] D. Perrin, An introduction to finite automata on infinite words, *in Automata on Infinite Words* (M. Nivat and D. Perrin, Eds.), Lecture Notes in Computer Science **192**, 1985, 2-17.

[23] D. Perrin and J.E. Pin, First order logic and star-free sets, *J. Comput. System Sci.*. **32**, 1986, 393-406.

[24] D. Perrin and P.E. Schupp, Automata on the integers, recurrence distinguishability, and the equivalence and decidability of monadic theories, *Proc. 1st Symp. on Logic in Computer Sci.*, 1986, 301-304.

[25] J.E. Pin, *Varieties of formal languages* , North Oxford Academic, London and Plenum, New-York, 1986.

[26] D. Thérien, A. Weiss, Graph congruences and wreath products, J. Pure Applied Algebra 35 (1985) 205-215.

[27] W. Thomas, The theory of successor with an extra predicate, *Math. Ann.* **237** 1978, 121-132.

[28] W. Thomas, Star-free regular sets of ω-sequences, *Information and Control* **42**, 1979, 148-156.

[29] W. Thomas, A combinatorial approach to the theory of ω-automata, *Information and Control* **48**, 1981, 261-283.

[30] W. Thomas, Classifying regular events in symbolic logic, *Journal of Computer and System Sciences* **25**, 1982, 360-376.

Danièle Beauquier, J.E. Pin,
LITP, Tour 55-65, Université Paris VI,
4 Place Jussieu, 75252 Paris Cedex 05, FRANCE.