Languages and scanners

Danièle Beauquier and Jean-Eric Pin *

In this paper, we are interested in a special class of automata, called scanners. These machines can be considered as a model for computations that require only "local" information. Informally, a scanner is an automaton equipped with a finite memory and a "sliding" window of a fixed length. In a typical computation, the sliding window is moved from left to right on the input, so that the scanner can remember the factors of length smaller than or equal to the size of the window. In view of these factors, the scanner decides whether or not the input is accepted or rejected.

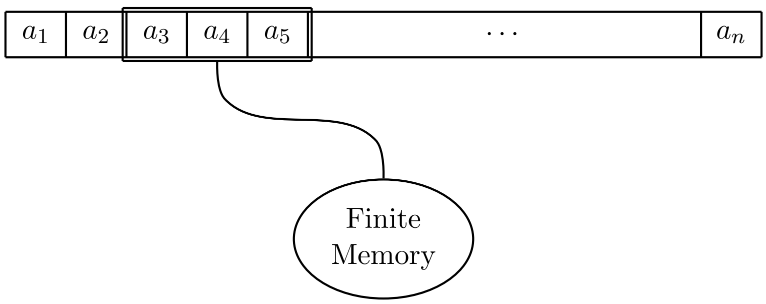


Figure 0.1: A scanner.

Scanners 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, a local language is described by the factors of length 2 of its words. 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 exactly $\{ca, ab, ba, bd\}$. The locally testable languages generalize local languages : the membership of a given word in such a language is determined by the set of factors of a fixed length k (the order in which these factors occur and their frequency is not relevant) of the word, and by the prefixes and suffixes of length $< k$ of the word. These conditions can be tested by a scanner. The locally testable languages are character-

*LITP/IBP, Université Paris VI et CNRS, Tour 55-56, 2 place Jussieu, 75251 Paris Cedex 05, FRANCE. **e-mail:** `beauquier@litp.ibp.fr, pin@litp.ibp.fr`

ized by a deep algebraic property of their syntactic semigroup, discovered independently by Brzozowski-Simon [2] and McNaughton [7].

There are several possible variations on this definition. First, one can drop the conditions about the prefixes and suffixes. Membership in this type of language, that we call strongly locally testable (SLT), is determined only by factors of a fixed length k. Thus, a language is SLT if and only if it is a finite boolean combination of languages of the form A^*wA^*, where w is a word. Surprisingly, this rather natural family of 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. But this time, the syntactic semigroup does not suffice, and a property of the image of the language in its syntactic semigroup is needed.

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 can only 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 $t \geq n$. This defines the threshold locally testable languages (TLT). A combination of two deep results of Straubing [11] and Thérien and Weiss [13] yields a syntactic characterization of these languages. In view of the results of the previous paragraph, it is reasonable to think that similar results hold if one drops the conditions about the prefixes and suffixes. However, we are not yet able to solve this problem.

The families of languages we have introduced are also deeply connected with the study of first order theory of successor, with a predicate for each letter, interpreted on finite words. Indeed Thomas [14] proved that the languages definable in this logic are exactly the TLT languages. Since we have an effective syntactic characterization of these languages, we derive the following decidability result : given a monadic second order formula φ of the theory of successor, it is decidable whether φ is equivalent to a first order formula. We also show that the languages definable by a boolean combination of existential formulas are exactly the SLTT languages.

Finally, scanners can also be used to define sets of infinite (or even bi-infinite) words. This is technically more difficult and will be the subject of a future paper. Most of the results of the present paper as well as those of this future paper have been presented, without proofs, at the ICALP in Stresa, 1989 [1].

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 up_k and us_k, respectively, the prefix and suffix of length k of u. 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. For instance $\begin{bmatrix} abababa \\ aba \end{bmatrix} = 3$, since aba occurs in three different places in $abababa$: $\underline{aba}baba$, $ab\underline{aba}ba$, $abab\underline{aba}$.

1.2 Finite semigroups.

Recall that a semigroup is a set S equipped with an associative multiplication. All semigroups considered in this paper are finite, except for free semigroups and free monoids. Therefore the word "semigroup" will mean "finite semigroup" in the rest of the paper. An element e of a semigroup S is idempotent if $e^2 = e$. The set of idempotents of a semigroup S is denoted by $E(S)$. Every non-empty finite semigroup contains at least one idempotent. If $S = E(S)$, S is an idempotent semigroup.

A monoid is a semigroup with an identity. Let S be a semigroup. We denote by S^1 the monoid equal to S if S has an identity, and to $S \cup \{1\}$, where 1 is a new identity, otherwise.

Given two semigroups S and T, a semigroup morphism $\varphi : S \to T$ is a function from S into T such that, for every $s, s' \in S, (s\varphi)(s'\varphi) = (ss')\varphi$. Recall the definitions of the Green's relations $\mathcal{R}$, $\mathcal{L}$ and $\mathcal{D}$: $s \mathrel{\mathcal{R}} t$ if and only if there exists $u, v \in S^1$ such that $su = t$ and $tv = s$, $s \mathrel{\mathcal{L}} t$ if and only if there exists $u, v \in S^1$ such that $us = t$ and $vt = s$, $s \mathrel{\mathcal{D}} t$ if and only if there exists $u \in S^1$ such that $s \mathrel{\mathcal{R}} u$ and $u \mathrel{\mathcal{L}} t$. We denote by $\leq_{\mathcal{J}}$ the preorder on S defined by $s \leq_{\mathcal{J}} t$ if and only if there exists $u, v \in S^1$ such that $usv = t$. One can show that $s \mathrel{\mathcal{D}} t$ if and only if $s \leq_{\mathcal{J}} t$ and $t \leq_{\mathcal{J}} s$. We shall need the following technical result.

Lemma 1.1 *Let S and T be two finite semigroups, let $\pi : S \to T$ be a surjective morphism. Let t and t' be two elements of T such that $t \mathrel{\mathcal{R}} t'$ (respectively $t \mathrel{\mathcal{L}} t'$, $t \mathrel{\mathcal{D}} t'$). Then there exist $s, s' \in S$ such that $s\pi = t$, $s'\pi = t'$ and $s \mathrel{\mathcal{R}} s'$ (respectively $s \mathrel{\mathcal{L}} s'$, $s \mathrel{\mathcal{D}} s'$).*

Proof. We give the proof for $\mathcal{R}$ only, but the proof for $\mathcal{L}$ and $\mathcal{D}$ is similar. Let t and t' be two elements of T such that $t \mathrel{\mathcal{R}} t'$. Since π is surjective, the set $t\pi^{-1}$ is non empty. Choose a minimal element s (for the preorder $\leq_{\mathcal{J}}$) in $t\pi^{-1}$. Since $t \mathrel{\mathcal{R}} t'$, there exist $u, v \in T^1$ such that $t' = tu$ and $t = t'v$. Let $x, y \in S^1$ be such that $x\pi = u$ and $y\pi = v$. Then $(sxy)\pi = tuv = t'v = t$ and $sxy \leq_{\mathcal{J}} sx \leq_{\mathcal{J}} s$. Thus, by the choice of s, we have $sxy \mathrel{\mathcal{D}} s$. It follows, by a standard argument ([9], proposition 1.4, p.47), $sxy\ \mathcal{R}\ s$, whence $s' = sx\ \mathcal{R}\ s$. This proves the lemma, since $s'\pi = tu = t'$. $\square$

Let S be a finite semigroup. A local subsemigroup of S is a subsemigroup of S of the form eSe, where $e \in E(S)$. A semigroup is said to be locally trivial, (respectively locally commutative, locally idempotent, locally a group, etc.) if all the local subsemigroups of S are trivial (respectively commutative, idempotent, groups, etc.). For instance, a semigroup S is locally idempotent and commutative if, for each $e \in E(S)$ and each $s, t \in S$, $(ese)^2 = (ese)$ and $(ese)(ete) = (ete)(ese)$.

1.3 Semidirect products.

Let M be a monoid and let T be semigroup. We write the product in M additively to provide a more transparent notation , but it is not meant to suggest that M is commutative. A left action of T on M is a map $(t, m) \rightarrow tm$ from $T \times M$ into M such that, for all $m, m_1, m_2 \in M$ and $t, t_1, t_2 \in T$, $(t_1t_2)m = t_1(t_2m)$ and $t(m_1 + m_2) = tm_1 + tm_2$, $t0 = 0$ Given such a left action, the semidirect product of M and T (with respect to this action) is the semigroup $M * T$ defined on the set $M \times T$ by the product $(m_1, t_1)(m_2, t_2) = (m_1 + t_1m_2, t_1t_2)$.

1.4 Counting semirings.

We define the following congruences on $\mathbb{N}$. $x \equiv y$ threshold t (also denoted $x \equiv_t y$) if and only if ($x < t$ and $x = y$) or ($x \geq t$ and $y \geq t$). For instance the equivalence classes of $\equiv_4$ are $\{0\}$, $\{1\}$, $\{2\}$, $\{3\}$, $\{4, 5, 6, 7, \ldots\}$. The quotient semiring $\mathbb{N}_t = \mathbb{N}/{\equiv_t}$ is called a counting semiring. In particular, the boolean semiring $\mathbb{B} = \mathbb{N}_1$ is a counting semiring.

2 Scanners and languages defined by factors of words.

2.1 Some equivalences and congruences.

Factors of words can be used in many different ways to define families of languages. We have selected four of them, which form the subject of this article. For every $k, t > 0$, let $\equiv_{k,t}$ be the equivalence of finite index defined on A^+ by setting $u \equiv_{k,t} v$ if and only if, for every word x of length $\leq k$, $\binom{u}{x} \equiv_t \binom{v}{x}$. For instance, $u \equiv_{k,1} v$ if and only if u and v have the same sets of factors of length k, and $u \equiv_{k,5} v$ if and only if u and v have the same factors of length k, but counted with multiplicity threshold 5.

Example 2.1 $abababab \equiv_{2,3} abababa$ since $abababab$ contains 4 ($\equiv$ 3 threshold 3) occurrences of ab and 3 ($\equiv$ 3 threshold 3) occurrences of ba, and no occurrences of aa (respectively bb).

We also define a congruence $\sim_{k,t}$ of finite index on A^+ by setting $u \sim_{k,t} v$ if

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

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

(c) $u \equiv_{k,t} v$.

These four equivalences define four classes of languages. A subset of A^+ is *locally k-testable* if it is union of $\sim_{k,1}$-classes. It is *threshold locally k-testable* if it is union of $\sim_{k,t}$-classes for some t. If one replaces in the previous definitions the congruence $\sim_{k,t}$ by the equivalence $\equiv_{k,t}$, one defines the corresponding notions of *strongly locally k-testable* and *strongly threshold locally k-testable* languages.

A subset of A^+ is *locally testable* (LT) if it is locally k-testable for some $k > 0$. The notions of *threshold locally testable*, *strongly locally testable*, *strongly threshold locally testable* languages are defined similarly. The corresponding abreviations are TLT, SLT, and STLT.

2.2 A combinatorial description.

One can also define these classes as boolean algebras. Set, for $x \in A^+$, and $r, t \geq 0$,

$$L(x, r, t) = \{u \in A^+ \mid \begin{bmatrix} u \\ x \end{bmatrix} \equiv_t r\}$$

Thus $L(x, r, t)$ is the set of all words u containing r occurrences of the factor x, but counted threshold t. For instance, $L(x, 1, 1) = A^*xA^*$, and $L(x, 0, 1) = A^+ \setminus A^*xA^*$.

Proposition 2.1 *Let L be a subset of A^+.*

(1) *L is SLT if and only if it is a boolean combination of languages of the form $L(x, 1, 1)$,*

(2) *L is STLT if and only if it is a boolean combination of languages of the form $L(x, r, t)$,*

(3) *L is LT if and only if it is a boolean combination of languages of the form $L(x, 1, 1)$, xA^*, or A^*x,*

(4) *L is TLT if and only if it is a boolean combination of languages of the form $L(x, r, t)$, xA^*, or A^*x.*

Proof. We only recall the proof of (1), which is standard. The proof of the other statements can be easily adapted. For each $u \in A^+$, put

$$E_k(u) = \{x \in A^+ \mid x \text{ is a factor of length } k \text{ of } u\} \text{ and } F_k(u) = A^k \setminus E_k(u)$$

Then the equivalence class of u with respect to $\equiv_{k,1}$ is the set

$$S(u) = \bigcap_{x \in E_k(u)} A^*xA^* \setminus \bigcup_{x \in F_k(u)} A^*xA^*$$

This shows that $S(u)$ is a boolean combination of languages of the form A^*xA^*.

Conversely, let L be a finite boolean combination of languages of the form A^*xA^*. Let k be the maximal length of the words x occurring in such a boolean expression. It suffices to show that if $|x| \leq k$, then A^*xA^* is a union of $\equiv_{k,1}$-classes. But this is clear, since if $u \in A^*xA^*$ and $u \equiv_{k,1} v$, then u and v have the same factors of length $|x|$, whence $v \in A^*xA^*$. $\square$

The relations between the four classes is shown in the following diagram.

$$\begin{tikzcd}
 & STLT \arrow[dr] & \\
SLT \arrow[ur] \arrow[dr] & & TLT \\
 & LT \arrow[ur] &
\end{tikzcd}$$

Example 2.2 Let $A = \{a, b\}$. Then $(ab)^+$ is locally testable since $(ab)^+ = (aA^* \cap A^*b) \setminus (A^*aaA^* \cup A^*bbA^*)$. We shall see in section 3 that $(ab)^+$ is not strongly locally testable.

Example 2.3 Let $A = \{a, b\}$. Then the set a^*ba^* is strongly threshold locally testable : it is the set of all words containing exactly one occurrence of b. We shall see in section 3 that a^*ba^* is not locally testable.

2.3 Scanners.

Our four classes of languages can also be defined in terms of a special class of finite automata, the scanners. The informal definition of a scanner has been given in the introduction. There are actually two types of scanners, depending on the use of the window : normally, a window of size k is allowed to read only factors of length k. If the scanner is unbounded, we allow the window to be also moved beyond the first and the last letter of the word, so that the prefixes and suffixes of length $< k$ can be read. For instance, if $k = 3$, and $u = abbaaabab$, different positions of the window are represented in the following diagrams :

$\boxed{a}\,bbaaabab \quad \boxed{ab}\,baaabab \quad \boxed{abb}\,aaabab \quad a\boxed{bba}\,aabab \quad \cdots \quad abbaaaba\boxed{b}$

We now give the formal definition. A scanner on a counting semiring K is a triple $S = (A, k, F)$ where A is a finite set (the alphabet), k is a positive integer (the size of the window), F is a (finite) set of polynomials of $K\langle A^*\rangle$ of degree k, called the memory. Let $u \in A^+$ be a word. Let $f : A^+ \to K\langle A^*\rangle$ be the application defined by

$$u\varphi = \sum_{|x|=k} \begin{bmatrix} u \\ x \end{bmatrix} x$$

Thus $u\varphi$ is just the formal sum of all factors of length k of u, with multiplicity counted in K. For instance, if $A = \{a, b\}$, $K = \mathbb{N}_3$, and $k = 3$,

$$\begin{aligned}(abbabbaabababbabbab)\varphi &= 4abb + 4bba + 5bab + baa + aab + 2aba \\ &= 3abb + 3bba + 3bab + baa + aab + 2aba\end{aligned}$$

A word u is accepted by S if and only if $uf \in F$.

A scanner on the boolean semiring is a *boolean scanner*. In the case of an unbounded scanner, the window is allowed to read the prefixes and suffixes of length $< k$. To represent this information, one introduces two new functions $\pi, \sigma : A^+ \to K\langle A^*\rangle$ defined by $u\pi = \sum_{t<k} up_t$ and $u\sigma = \sum_{t<k} us_t$. The memory is now a triple (P, F, S) of sets of polynomials of $K\langle A^*\rangle$ of degree $\leq k$. Intuitively, P codes the set of possible prefixes, F the set of possible factors, and S the set of possible suffixes. A word u is accepted by S if and only if $u\pi \in P$, $u\varphi \in F$ and $u\sigma \in S$.

Proposition 2.2 *Let L be a subset of A^+.*

(1) *L is SLT if and only if it is accepted by a boolean scanner,*

(2) *L is STLT if and only if it is accepted by a scanner,*

(3) *L is LT if and only if it is accepted by an unbounded boolean scanner,*

(4) *L is TLT if and only if it is accepted by an unbounded scanner,*

Proof. Again, we only give the proof for (1), but the other proofs are similar. Let $S = (A, k, F)$ be a boolean scanner recognizing a subset L of A^+. Observe that $u \equiv_{k,1} v$ if and only if $uf = vf$. It follows that L is union of $\equiv_{k,1}$-classes. Conversely, if L is union of $\equiv_{k,1}$-classes for some k, put $F = uf \mid u \in L$. Then the boolean scanner $S = (A, k, F)$ recognizes L. $\square$

3 Syntactic characterizations.

In this section, we give effective characterizations for three of the four families of languages introduced above. In order to keep a uniform notation for the subsequent statements, we shall denote by $S(L)$ the syntactic semigroup of a recognizable language L of A^+, by $\eta : A^+ \to S(L)$ its syntactic morphism, and by $P(L) = L\eta$ the syntactic image of L. We need first to introduce some algebraic tools.

3.1 Varieties of semigroups.

A variety of (finite) semigroups is a class of semigroups closed under taking subsemigroups, morphic images (or quotients) and finite direct products. Varieties of monoids are defined similarly. The following varieties will be used in this paper.

- $\mathbf{J_1}$, the variety of all idempotent and commutative monoids,
- **Com**, the variety of all commutative monoids,
- **Acom**, the variety of all commutative aperiodic monoids,
- **LI**, the variety of locally trivial semigroups,
- $\mathbf{LI}_k$, the variety of all semigroups S that satisfy the equation

  $$x_1x_2\cdots x_kxx_1x_2\cdots x_k = x_1x_2\cdots x_k$$

  for all $x, x_1, x_2, \ldots, x_k \in S$.
- $\mathbf{LJ_1}$, the variety of locally idempotent and commutative semigroups,

Given a variety of monoids $\mathbf{V}$ and a variety of semigroups $\mathbf{W}$, we denote by $\mathbf{V} * \mathbf{W}$ the variety of semigroups generated by all the semidirect products of the form $M * T$, where $M \in V$ and $T \in W$.

3.2 Varieties of languages.

Let $\mathbf{V}$ be a variety of semigroups (monoids). One associates to each alphabet A the set $A^+\mathcal{V}$ $(A^*\mathcal{V})$ of all languages of A^+ (A^*) whose syntactic semigroup (monoid) belongs to $\mathbf{V}$. $\mathcal{V}$ is called the variety of languages corresponding to $\mathbf{V}$. A description of various varieties of languages can be found in the litterature [5, 6, 9].

Proposition 3.1 *For each alphabet A,*

(1) $A^+\mathcal{J}_1$ *is the boolean algebra generated by the languages of the form* A^*aA^* *where* $a \in A$,

(2) $A^+\mathcal{LI}_k$ *is the boolean algebra generated by the languages of the form* A^*u, vA^* *where u and v are words of length* $\leq k$,

(3) $A^+\mathcal{LI}$ *is the boolean algebra generated by the languages of the form* A^*u, vA^* *where* $u, v \in A^+$,

(4) *$A^*\mathcal{A}com$ is the boolean algebra generated by the languages of the form $\{u \in A^* \mid |u|_a = r\}$, where $a \in A$ and $r \in \mathbb{N}$.*

For now, we need a description, due to Straubing [11], of the varieties of languages corresponding to $\mathbf{V} * \mathbf{LI}$ and to $\mathbf{V} * \mathbf{LI}_k$, when V is a variety of monoids.

Let k be an integer, and let $B_k = A^k$. To avoid ambiguity, words of B_k^* will be represented by finite sequences $(b_1, b_2, \cdots, b_n)$, where each $b_i \in A^k$. We define a (sequential) function $\sigma_k : A^+ \to B_k^*$ by setting

$$w\sigma_k = 1 \text{ if } |w| < k,$$
$$(wa)\sigma_k = (w\sigma_k, (wa)s_k) \text{ if } |w| \geq k \text{ and } a \in A.$$

For example, $(abbaab)\sigma_3 = (abb, bba, baa, aab)$. Thus σ_k associates to a word u the sequence of words appearing on a window of size k when u is read from left to right.

To each congruence α of finite index on B_k^*, associate the congruence (α, k) on A^+ defined by $u(\alpha, k)v$ if and only if

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

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

(c) $u\sigma_k \alpha v\sigma_k$

Denote by $\mathcal{V}$ and $\mathcal{W}_k$ the varieties of languages corresponding to $\mathbf{V}$ and $\mathbf{V} * \mathbf{LI}_k$, respectively.

Theorem 3.2 [11] *A language belongs to $A^+\mathcal{W}_{k-1}$ if and only if it is a finite union of (α, k)-classes for some congruence α on B_k^* such that $B_k^*/\alpha \in V$.*

We give an equivalent form of Theorem 3.2 in terms of boolean algebras.

Corollary 3.3 *For every alphabet A, $A^+\mathcal{W}_{k-1}$ is the boolean algebra generated by the languages of the form A^*u, vA^* (where u and v are words of length $< k$) or $X\sigma_k^{-1}$, where $X \in B_k^*\mathcal{V}$.*

Proof. In one direction, it suffices to show that each of the languages A^*u, vA^* and $X\sigma_k^{-1}$ belong to $A^+\mathcal{W}_{k-1}$. First, if $|u| < k$, then, by proposition 3.1, $S(A^*u), S(uA^*) \in \mathbf{LI}_{k-1} \subset \mathbf{V} * \mathbf{LI}_{k-1}$. Furthermore, since σ_k is a sequential function, a general result ([5], Chap. 6) states that $S(X\sigma_k^{-1})$ divides a wreath product of the form $M(X) \circ S(\sigma_k)$, where $S(\sigma_k)$ is the syntactic invariant of σ_k. Now $M(X) \in \mathbf{V}$ since $X \in B_k^*\mathcal{V}$, and $S(\sigma_k) \in \mathbf{LI}_{k-1}$. It follows that $S(X\sigma_k^{-1}) \in \mathbf{V} * \mathbf{LI}_{k-1}$.

Conversely, if $L \in A^+\mathcal{W}_{k-1}$, then by Theorem 2.5, L is union of (α, k)-classes for a certain congruence α such that $B_k^*/\alpha \in \mathbf{V}$. Now, it follows from the definition of (α, k) that the equivalence classes of (α, k) are boolean combinations of sets of the form A^*u, vA^* (where u and v are words of length

$< k$) or $X\sigma_k^{-1}$, where X is an equivalence class for α. But since $B_k^*/\alpha \in \mathbf{V}$, one has $X \in B_k^*\mathcal{V}$. $\square$

For the variety $\mathbf{V} * \mathbf{LI}$, one has the following result.

Theorem 3.4 *Let S be a semigroup. Then $S \in \mathbf{V} * \mathbf{LI}$ if and only if $S \in \mathbf{V} * \mathbf{LI}_k$, where $k = |S|$.*

Applying these results with $\mathbf{V} = \mathbf{Acom}$ and $\mathbf{J_1}$ respectively, one obtains the following corollary.

Corollary 3.5

(1) *L is locally testable if and only if $S(L)$ belongs to $\mathbf{J_1} * \mathbf{LI}$,*

(2) *L is threshold locally testable if and only if $S(L)$ belongs to $\mathbf{Acom}*\mathbf{LI}$,*

Proof. We only prove (2), the proof of (1) beeing similar. Let L be a subset of A^+. By Corollary 3.3 and Theorem 3.4, $S(L) \in \mathbf{Acom}*\mathbf{LI}$ if and only if L is a boolean combination of languages of the form A^*u, vA^* or $X\sigma_k^{-1}$, where $X \in B_k^*\mathbf{Acom}$. Furthermore, by Proposition 3.1, $X \in B_k^*\mathbf{Acom}$ if and only if X is a boolean combination of languages of the form $\{u \in B_k^* \mid |u|_x = r\}$, where $x \in B_k$ and $r \in \mathbb{N}$. Now we have

$$\{u \in A^* \mid |u|_x = r\}\sigma_k^{-1} = \{u \in A^* \mid \begin{bmatrix} u \\ x \end{bmatrix} = r\} = L(x, r, t) \text{ for every } t > r.$$

Therefore $S(L) \in \mathbf{Acom} * \mathbf{LI}$ if and only if L is a boolean combination of languages of the form A^*u, vA^* or $L(x, r, t)$, that is, if and only if L is threshold locally testable. $\square$

Note that corollary 3.5 does not give an algorithm to decide whether a language is LT, TLT or PLT. Indeed, it is not clear at first sight whether one can decide if a given semigroup belongs to $\mathbf{J_1} * \mathbf{LI}$ or $\mathbf{Acom} * \mathbf{LI}$. But Brzozowski and Simon [2] and Mac Naughton [7] shaw independently that $\mathbf{J_1} * \mathbf{LI} = \mathbf{LJ_1}$. Therefore we have

Theorem 3.6 [2, 7, 11] *L is locally testable if and only if $S(L)$ belongs to $\mathbf{LJ_1}$.*

The syntactic characterization of locally threshold testable languages is more involved and depends on a deep result of Thérien and Weiss [13]. Given a semigroup S, 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.

Theorem 3.7 [13] *A semigroup S belongs to $\mathbf{Acom}*\mathbf{LI}$ if and only if S is aperiodic and its graph satisfies the following condition (C): if p and r are edges from e to f, and if q is an edge from f to e, then $pqr = rqp$.*

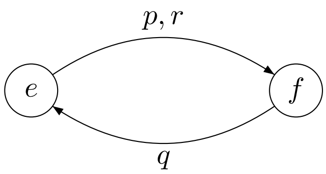


Figure 3.2: The condition $pqr = rqp$.

Therefore, one obtains

Corollary 3.8 *A language L is threshold locally testable if and only if $S(L)$ is aperiodic and its graph satisfies condition (C).*

Example 3.1 Let $A = \{a, b\}$, and let $L = a^*ba^*$. Then L is recognized by the automaton

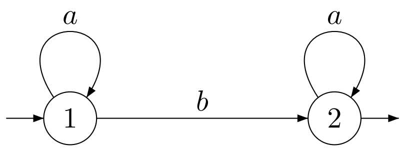


Figure 3.3: The minimal automaton of a^*ba^*.

The transitions are given in the following table

	1	2
a	1	2
b	2	–
bb	–	–

Therefore, $S = S(L)$ is presented by the relations $a = 1$, $b^2 = 0$. Thus $S = \{a, b, 0\}$, where $a = 1$ is the identity, and $E(S) = \{1, 0\}$. The local semigroups are $0S0 = \{0\}$, and $1S^1 = S$. This last local semigroup is not idempotent, since $b^2 \neq b$. Therefore, L is not locally testable. On the other hand, $G(S)$ is the graph represented below, which satisfies condition (C).

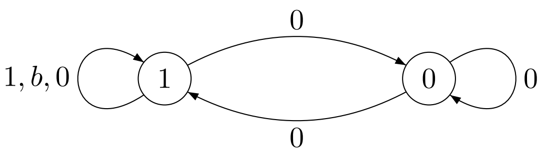


Figure 3.4: The graph of S.

Therefore L is TLT (see example 2.2).

The three classes of languages we have considered so far were characterized by an algebraic property of their syntactic semigroup. Such a property do not suffice, however, to characterize the class of strongly locally testable languages. To overcome this difficulty, we need to consider not only the syntactic semigroup, but also the syntactic image of the language.

Let S be a 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 $D \subset P$. It is equivalent to say that $s \in P$ and $s \mathrel{\mathcal{D}} t$ imply $t \in P$. The next proposition shows that this property is stable under quotients.

Proposition 3.9 *Let S and T be two semigroups, let $\pi : S \to T$ be a surjective morphism. Let P_S (respectively P_T) be a subset of S (respectively T) such that $P_T = P_S\pi$ and $P_S = P_T\pi^{-1}$. Then P_S saturates the $\mathcal{D}$-classes of S if and only if P_T saturates the $\mathcal{D}$-classes of T.*

Proof. Suppose that P_T saturates the $\mathcal{D}$-classes of T. Let $s \in P_S$ and $s' \in S$ such that $s \mathrel{\mathcal{D}} s'$. Then $s\pi \mathrel{\mathcal{D}} s'\pi$, $s\pi \in P_T$ and therefore, $s'\pi \in P_T$. Thus $s' \in P_T\pi^{-1} = P_S$, and P_S saturates the $\mathcal{D}$-classes of S.

Conversely, suppose that P_S saturates the $\mathcal{D}$-classes of S. Let $t, t' \in T$ such that $t \in P_T$ and $t \mathrel{\mathcal{D}} t'$. By Lemma 1.1, there exists $s, t \in S$ such that $s\pi = t$, $s'\pi = t'$ and $s \mathrel{\mathcal{D}} s'$. It follows that $s \in P_T\pi^{-1} = P_S$, whence $s' \in P_S$ and $s'\pi = t' \in P_S\pi = P_T$. $\square$

Theorem 3.10 *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 $S(L)$.*

Proof. To simplify notations, put $S = S(L)$, $P = P(L)$ and denote by $\sim_k$ the congruence $\sim_{k,1}$. If L is SLT, then L is also LT and thus S is locally idempotent and commutative by Theorem 3.6. Furthermore, L is a boolean combination of languages of the form A^*xA^*. Let k be the maximal length of the words x occurring in this boolean combination. Then, if $u \in L$ and if u and v have the same factors of length $\leq k$, then $v \in L$. We claim that P saturates the $\mathcal{R}$-classes of S. Let $s \in P$ and let $t \in S$ such that $s \mathrel{\mathcal{R}} t$. Then there exist some elements $x, y \in S(L)^1$ such that $sx = t$ and $ty = s$. Let $s' \in A^+$, $x', y' \in A^*$ be words such that $s'\eta = s$, $x'\eta = x$ and $y'\eta = y$ (if $x = 1$ or $y = 1$, we take $x' = 1$ or $y' = 1$, respectively). Now the word $s'(x'y')^k$ belongs to L, since $(s'(x'y')^k)\eta = s(xy)^k = s$. Furthermore, the words $s'(x'y')^k$ and $s'(x'y')^kx'$ contain the same factors of length $\leq k$. This is obvious if $x' = 1$. Suppose now $x' \in A^+$. Then every factor of $s'(x'y')^k$ is clearly a factor of $s'(x'y')^kx'$. Conversely, let t be a factor of length $\leq k$ of $s'(x'y')^kx'$. Then either t is a factor of $s'(x'y')^k$, or t is a factor of $(x'y')^{k-1}x'$, since $|(x'y')^{k-1}| \geq k-1$. But $(x'y')^{k-1}x'$ itself is a factor of $s'(x'y')^k$ and thus t is a factor of $s'(x'y')^k$. It follows, by the remark above, that $s'(x'y')^kx'$ belongs to L and hence $(s'(x'y')^kx')\eta = s(xy)^kx = sx = t \in P$, proving the claim. A dual argument would show that P saturates the $\mathcal{L}$-classes, and hence P also saturates the $\mathcal{D}$-classes.

Conversely, assume that S is locally idempotent and commutative and that P saturates the $\mathcal{D}$-classes of S. Then, L is locally testable by theorem 3.6, and thus is union of $\sim_k$-classes for some k. Put $S_k = A^+/{\sim_k}$. Then there is a surjective morphism $\pi_k : S_k \to S$, and a subset Q of S_k such that

(a) $L = Q\pi_k^{-1}$ and $Q = L\pi_k$,

(b) $Q = P\pi^{-1}$ and $P = Q\pi$.

Now, by Proposition 3.9, Q saturates the $\mathcal{D}$-classes of S_k. To finish the proof, we need a result on the $\mathcal{D}$-classes of S_k. Denote by $F_k(u)$ the set of factors of length k of a word u.

Lemma 3.11 *Let u and v be two words of A^+. Then $u\pi_k \ \mathcal{D}\ v\pi_k$ if and only if either $u = v$ or $F_k(u) = F_k(v) \neq \emptyset$ (this case implies that u and v are of length $\geq k$).*

Proof. We first treate the case $|u| < k$ (respectively $|v| < k$). If $u\pi_k \ \mathcal{D}\ v\pi_k$, then there exist four words $x, y, s, t \in A^*$ such that $xuy \sim_k v$ and $svt \sim_k u$. This implies $svt = u$, whence $|v| < k$, and thus $xuy = v$, so that $u = v$.

Suppose now $|u|, |v| \geq k$. If $u\pi_k \ \mathcal{D}\ v\pi_k$, there exist two words $x, y \in A^*$ such that $xuy \sim_k v$. In particular, every factor of length k of u is a factor of v, and, by a dual argument, $F_k(u) = F_k(v)$.

Conversely, assume that $F_k(u) = F_k(v)$ and let p (respectively s) be the prefix (suffix) of length k of u. Then p (respectively s) occurs in v, so that $v = v_0pv_1 = v_2sv_3$. Put $w = v_0uv_3$. We claim that $w \sim_k v$. Indeed v_0p (respectively sv_3) is a common prefix (suffix) of v and w. Next, since $F_k(u) = F_k(v)$, each factor of length k of v is a factor of u and hence a factor of w. Conversely, let t be a factor of length k of w. Then t is either a factor of v_0p, a factor of u, or a factor of sv_3. In each case, it is also a factor of v, which proves the claim. Therefore $v\pi_k = (v_0\pi_k)(u\pi_k)(v_3\pi_k)$, and, by a dual argument, $u\pi_k \ \mathcal{D}\ v\pi_k$. $\square$

We now conclude the proof of Theorem 3.10. We start with the equality $L = Q\pi_k^{-1}$ and we distinguish two categories of elements in Q. Put

$$Q_1 = \{s \in Q \mid \text{ every word of } s\pi_k^{-1} \text{ is of length } \geq k\} \text{ and}$$
$$Q_2 = \{s \in Q \mid \text{ there exists a word of } s\pi_k^{-1} \text{ of length } < k\}$$

Then, since $Q = Q_1 \cup Q_2$, we have

$$L = Q_1\pi_k^{-1} \cup \bigcup_{s \in Q_2} s\pi_k^{-1}$$

and we shall prove separately that the languages $Q_1\pi_k^{-1}$ and $s\pi_k^{-1}$, for $s \in Q_2$, are SLT.

Let $s \in Q_2$. Then $s\pi_k^{-1}$ contains a word u of length $< k$, and $s\pi_k^{-1} = \{u\}$, since u cannot be equivalent to another word. But

$$\{u\} = A^*uA^* \setminus \left(\bigcup_{a \in A} A^*uaA^* \cup \bigcup_{a \in A} A^*auA^* \right)$$

and thus $s\pi_k^{-1}$ is strongly locally testable.

Since Q saturates the $\mathcal{D}$-classes of S_k, and since, by Lemma 3.11, an element of Q_2 cannot be $\mathcal{D}$-equivalent with an element of Q_1, Q_1 is a union of $\mathcal{D}$-classes. Furthermore, Lemma 3.11 shows that a $\mathcal{D}$-class $\mathcal{D}$ contained in Q_1 is entirely characterized by a certain non empty set F of words of length k. More precisely $D\pi_k^{-1} = \{u \in A^+ \mid F_k(u) = F\}$. It follows that $D\pi_k^{-1}$ is strongly locally testable, since

$$\{u \in A^+ \mid F_k(u) = F\} = \bigcap_{x \in F} A^*xA^* \setminus \bigcup x \in A^k \setminus FA^*xA^*$$

Finally, L is a finite union of SLT languages, and thus is also strongly locally testable. □

Example 3.2 Let $A = \{a, b, c\}$, and let $L = c(ab)^* \cup c(ab)^*a$. Then L is recognized by the following automaton

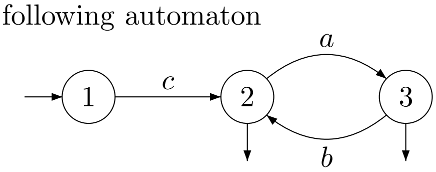


Figure 3.5: An automaton recognizing $L = c(ab)^* \cup c(ab)^*a$.

The transitions are given in the following table

	1	2	3
a	–	3	–
b	–	–	2
c	2	–	–
ab	–	2	–
ba	–	–	3
ca	3	–	–

Therefore, $S(L)$ is presented by the relations $cabc = c$, $a^2 = b^2 = c^2 = bc = ac = 0$. The $\mathcal{D}$-class structure is represented in the following diagram, where the grey box is the image of L.

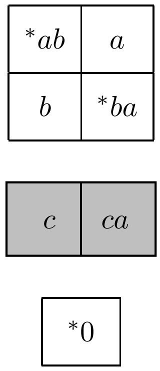


Figure 3.6: The $\mathcal{J}$-class structure.

Thus $P(L)$ saturates the $\mathcal{D}$-classes, and L is SLT. In fact,

$$L = A^*cA^* \setminus (A^*aaA^* \cup A^*acA^* \cup A^*bbA^* \cup A^*bcA^* \cup A^*cbA^* \cup A^*ccA^*).$$

The next statement summarizes the results of this section.

Corollary 3.12 *For a given recognizable subset L of A^+, the following properties are effectively decidable :*
(1) *L is locally testable,*
(2) *L is threshold locally testable,*
(3) *L is strongly locally testable.*

In view of these results, it is natural to conjecture that one can also decide whether a given language is STLT, but we don't have a proof of this fact.

4 Connections with logic.

The connections between formal languages and mathematical logic were first studied by Büchi [3]. To each word $u \in A^+$ is associated a structure

$$\mathcal{M}_u = (\{1, 2, \dots, |u|\}, S, (R_a)_{a \in A})$$

where S denotes the successor relation on $\{1, 2, \dots, |u|\}$ and R_a is set of all i such that the i-th letter of u is an a. For instance, if $A = \{a, b\}$ and $u = abaab$, then $R_a = \{1, 3, 4\}$ and $R_b = \{2, 5\}$. The logical language appropriate to such models has S and the R_a's as non logical symbols, and formulas are built in the standard way by using these non-logical symbols, variables, boolean connectives, equality and quantifiers. Note that we don't use the symbol $<$ in this logic. We shall denote by $\mathcal{L}_1(A)$ and $\mathcal{L}_2(A)$, respectively, the set of first order and monadic second order formulas of this logic. Given a sentence φ, we denote by $L(\varphi)$ the set of all words which satisfy φ, when φ is interpreted according to the model described above. For instance, if

$$\varphi = \exists x \exists y \exists z((y = Sx) \wedge (z = Sy) \wedge R_a x \wedge R_b y \wedge R_b z)$$

then $L(\varphi) = A^* abbA^*$.

The seminal result of Büchi can now be stated as follows

Theorem 4.1 [3] *A subset of A^+ is rational if and only if it can be defined by a $\mathcal{L}_2(A)$-sentence.*

The first order theory was investigated by Thomas [14]. Thomas proved that a language is TLT if and only if it is definable by a $\mathcal{L}'_1(A)$-sentence, where $\mathcal{L}'_1(A)$ is the logical language obtained by completing $\mathcal{L}_1(A)$ with the 0-ary symbols min, max, interpreted as the minimum 1 and the maximum $|u|$ on $\{1, \ldots, |u|\}$. Furthermore, Thomas proved that boolean combinations of existential $\mathcal{L}'_1(A)$-sentences were sufficient to define TLT-languages. Now, it is easy to define min and max in terms of S. For instance, one can consider min as a new variable satisfying the formula $\forall x \neg S(x, min)$. Therefore, one obtains

Theorem 4.2 *A subset of A^+ is threshold locally testable if and only if it is definable by a $\mathcal{L}_1(A)$-sentence.*

However, the situation is slightly different if one considers only boolean combinations of existential $\mathcal{L}_1(A)$-sentences, because rewriting min in terms of S creates some alternations of quantifiers. More precisely, we have

Theorem 4.3 *A subset of A^+ is strongly threshold locally testable if and only if it is definable by a boolean combination of existential $\mathcal{L}_1(A)$-sentences.*

Proof. If L is a STLT language, L is a boolean combination of languages of the form $L(x, r, t)$. Therefore it suffices to show that each of these languages can be defined by a boolean combination of existential $\mathcal{L}_1(A)$-sentences. The formal proof can easily be adapted from the following example, where $A = \{a, b\}$. One has $L(ab, 2, 3) = L(\varphi)$, where $\varphi = \varphi_1 \wedge \neg\varphi_2$, and

$$\begin{aligned}
\varphi_1 = {} & \exists x_1 \exists x_2 \exists x_3 \exists x_4 (\neg(x_1 = x_3) \wedge S(x_1, x_2) \wedge S(x_3, x_4) \\
& \wedge R_a x_1 \wedge R_b x_2 \wedge R_a x_3 \wedge R_b x_4) \\
\varphi_2 = {} & \exists x_1 \exists x_2 \exists x_3 \exists x_4 \exists x_5 \exists x_6 (\neg(x_1 = x_3) \wedge \neg(x_3 = x_5) \wedge \neg(x_1 = x_5) \\
& \wedge S(x_1, x_2) \wedge S(x_3, x_4) \wedge S(x_5, x_6) \\
& \wedge R_a x_1 \wedge R_b x_2 \wedge R_a x_3 \wedge R_b x_4 \wedge R_a x_5 \wedge R_b x_6)
\end{aligned}$$

Conversely, it suffices to show that a language L defined by an existential $\mathcal{L}_1(A)$-sentence φ is SLTT. We use an argument of game theory, which we borrow from [14]. For the conveniance of the reader, we briefly review the terminology of game theory needed to achieve the proof (see [10] for more details). Let $u = u_1 \cdots u_r$ and $v = v_1 \cdots v_s$ be two words, where $u_1, \ldots, u_r, v_1, \ldots, v_s \in A$. A position in u (respectively in v) is a an element of $\{1, \ldots, |u|\}$ (respectively $\{1, \ldots, |v|\}$). For each $m > 0$, define a game $G_m(u, v)$ between two players as follows. Player I plays first, and chooses m positions $i_1, \ldots, i_m$ of u. Then player II must choose m positions $j_1, \ldots, j_m$ in v. We say that II wins the game if the following conditions are satisfied :

(1) For $1 \leq r \leq m$, the letter u_{i_r} is equal to the letter v_{j_r}. (Intuitively, if I chooses an occurrence of a letter a, then II should choose an occurrence of the same letter.)

(2) For each $r, s \leq m$, $i_r = i_s$ if and only if $j_r = j_s$. (Intuitively, if I decides to choose twice - or more - the same position, then II should follow this choice. Conversely, II is not allowed to choose twice the same position if it was not the choice of I.)

(3) For each $r, s \leq m$, $i_r = i_s + 1$ if and only if $j_r = j_s + 1$. (Intuitively, if I decides to choose two adjacent positions, then II should follow this choice. Similarly, if I chooses two non-adjacent positions, then II should follow this choice.)

Now, by the theory of Ehrenfeucht-Fraïssé, two words u and v satisfy the same existential formulas of quantifier rank $\leq m$ if and only if Player II has a winning strategy in the games $G_m(u, v)$ and $G_m(v, u)$.

The main argument of the proof is the following lemma.

Lemma 4.4 *Let $n = 2^m + 1$ and $t = (m-1)(2^m + 1) + 1$. If $u \equiv_{n,t} v$, then Player II has a winning strategy in the games $G_m(u, v)$ and $G_m(v, u)$.*

Proof. By symmetry, it suffices to prove the result for the game $G_m(u, v)$. The strategy of II is easier to understand if one thinks that the choice of i_k (respectively j_k) also determines the segment

$$I_k = [1, \ldots, |u|] \cap [i_k - 2^{m-k}, i_k + 2^{m-k}]$$

(respectively $J_k = [1, \ldots, |v|] \cap [j_k - 2^{m-k}, j_k + 2^{m-k}]$). The strategy of II consists to choose j_k so that the following conditions are satisfied :

(a) $u[I_k] = v[J_k]$,

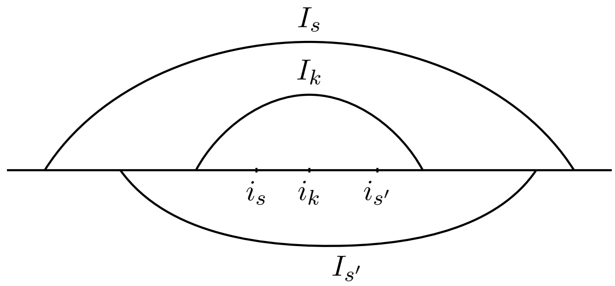
(b) For every $s < k$, $i_s \in I_k$ if and only if $j_s \in J_k$. In this case I_k is a subsegment of I_s and J_k is the corresponding subsegment of J_s.

We prove by induction on k that j_k can be choosen so that conditions (a) and (b) are satisfied. For $k = 1$, condition (b) is empty, and condition (a) can be fulfilled because u and v have the same factors of length n by hypothesis. Assume that j_s has been choosen successfully for $s < k$. We now choose j_k as follows.

First, assume there exists $s < k$ such that $i_s \in I_k$, and let us take the smallest s satisfying this condition. Then

$$i_s - 2^{m-s} \leq (i_k + 2^{m-k}) - 2^{m-s} \leq (i_k + 2^{m-k}) - 2.2^{m-k} = i_k - 2^{m-k}$$

and, symmetrically, $i_k + 2^{m-k} \leq i_s + 2^{m-s}$. Therefore I_k is a subsegment of I_s and we can take for J_k the corresponding subsegment of J_s. Since $u[I_s] = v[J_s]$, we have $u[I_k] = v[J_k]$. Now if $i'_s \in I_k$ for some s' such that $s < s' < k$, then $d(i_s, i_{s'}) \leq d(i_s, i_k) + d(i_k, i_{s'}) \leq 2.2^{m-k} \leq 2^{m-s'}$. Therefore $I_{s'}$ is a subsegment of I_s , as illustrated in the following diagram, and $J_{s'}$ is the corresponding subsegment of $J_{s'}$.

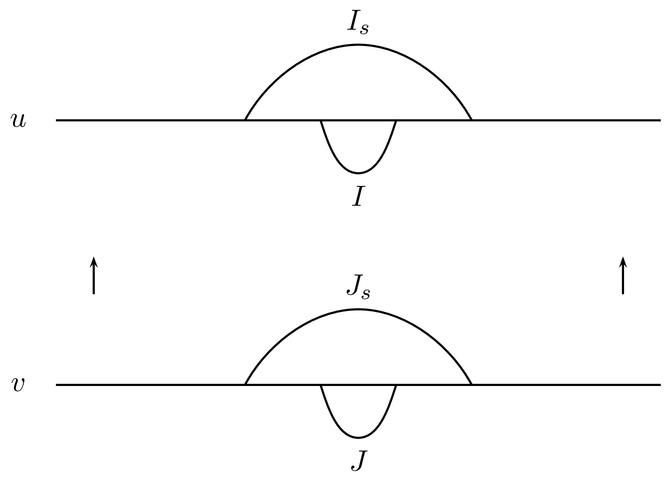


A similar argument would show that, if $j_{s'} \in J_k$, then $i_{s'} \in I_k$. Thus conditions (a) and (b) are satisfied in this case.

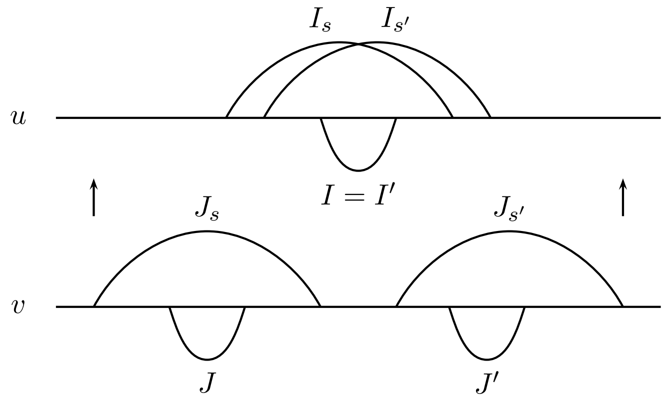
Now suppose that $i_s \in I_k$ for every $s < k$. We claim that there exists at least one occurrence of the factor $x = u[I_k]$ in v, defining a segment J_k such that$v[J_k] = x$ and $j_s \in J_k$ for every $s < k$. Indeed, assume that every segment J_k such that $v[J_k] = x$ satisfies $j_s \in J_k$ for some $s < k$, that is, $j_k \in [j_s - 2^{m-k}, j_s + 2^{m-k}]$. Then one can bound the number of occurrences of x in v as follows:

$$\begin{bmatrix} v \\ x \end{bmatrix} \leq \sum_{0<s<k} (2^{m-k+1} + 1)) = (k-1)(2^{m-k+1} + 1) < t$$

Now, since $|x| \leq n$, we have by hypothesis $\left[\begin{smallmatrix} u \\ x \end{smallmatrix}\right] \equiv \left[\begin{smallmatrix} v \\ x \end{smallmatrix}\right]$ threshold t, so that $\left[\begin{smallmatrix} u \\ x \end{smallmatrix}\right] = \left[\begin{smallmatrix} v \\ x \end{smallmatrix}\right]$. But each segment J in v such that $v[J] = x$ and $j_s \in J$ defines a segment I in u such that $u[I] = x$ and $i_s \in I$, since $u[I_s] = v[J_s]$ by the induction hypothesis.



Furthermore, this application is injective, that is, the situation represented in the figure below can never occur.



Indeed, we have seen above that if, for instance, $s < s'$, $I = I'$ implies that $I_{s'}$ is a subsegment of I_s. Therefore, $J_{s'}$ must be a subsegment of J_s and thus $J = J'$. It follows that each occurrence of x in v that is a factor of some $v[J_s]$ is in one-to-one correspondence with an occurrence of x in u that is a factor of $u[I_s]$. In particular, $\left[{u \atop x}\right] > \left[{v \atop x}\right]$, a contradiction. This proves the claim, and conditions (a) and (b) can be satisfied.

Now, it is easy to verify that this choice of $j_1, \ldots, j_m$ is a winning strategy for player II. $\square$

We now conclude the proof of theorem 4.3. Assume that L is defined by an existential sentence φ of quantifier rank m. If $u \equiv_{n,t} v$, then Lemma 4.4 and the theorem of Ehrenfeucht-Fraïssé shows that u satisfies φ if and only if v satisfies φ. This means that $L(\varphi)$ is a union of $\equiv_{n,t}$-classes, and hence $L(\varphi)$ is strongly locally threshold testable. $\square$

5 Remarks.

There are a few extensions that were not considered in order to keep this paper to a reasonable size. The first possibility would be to introduce modulo counting. If one considers modulo counting only, the notions of "periodically locally testable language" and "modular scanner" can be easily defined, and the syntactic characterization follows from the works of Straubing and Thérien (the condition would be that $S(L)$ is locally a commutative group). One can also give a logical interpretation if one allows the "modular" quantifiers considered by Straubing, Thérien and Thomas [12]. One can also consider simultaneously modulo and threshold counting. The corresponding variety of semigroups would be $\mathbf{Com} * \mathbf{LI}$, for which an effective description has been given by Thérien and Weiss [13] (a semigroup belongs to $\mathbf{Com} * \mathbf{LI}$ if and only if the graph associated with the semigroup satisfies (C)). However, no such decidability results is known for the corresponding "strong" notions. In conclusion, if one removes the conditions on prefixes and suffixes, nothing is known, except in the boolean case.

The second possible extension is to consider infinite words, and this will be the subject of a future paper.

References

[1] D. Beauquier and J.E. Pin, 1989, Factors of words, in *Automata, Languages and Programming*, (G. Ausiello, M. Dezani-Ciancaglini and S. Ronchi Della Rocca, eds.), Lecture Notes in Comput. Sci., **372**, Springer, 63–79.

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

[3] J.R. Büchi, 1962, On a decision method in restricted second-order arithmetic, in *Proc. 1960 Int. Congr. for Logic, Methodology and Philosophy of Science*, Stanford Univ. Press, Standford, 1–11.

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

[5] S. Eilenberg, 1976, *Automata, languages and machines*, Vol. B, Academic Press, New York.

[6] G. Lallement, 1979, *Semigroups and combinatorial applications*, Wiley, New York.

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

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

[9] J.-E. Pin, 1984 *Variétés de langages formels*, Masson, Paris; English translation: 1986, *Varieties of formal languages*, Plenum, New-York.

[10] J.G. Rosenstein, 1982, *Linear orderings*, Academic Press, New York.

[11] H. Straubing, Finite semigroup varieties of the form $V * D$, *J. Pure Applied Algebra* **36** (1985) 53–94.

[12] H. Straubing, D. Thérien and W. Thomas, 1988, Regular Languages Defined with Generalized Quantifiers, in *Proc. 15th ICALP*, Springer Lecture Notes in Computer Science **317**, 561–575.

[13] D. Thérien and A. Weiss, 1985, Graph congruences and wreath products, *J. Pure Applied Algebra* **35**, 205–215.

[14] W. Thomas, 1982, Classifying regular events in symbolic logic, *J. Comput. Syst. Sci* **25**, 360–375.