August 8, 2001

# Some results on the generalized star-height problem

Jean-Eric Pin*, H. Straubing†and D. Thérien‡

Jean-Eric.Pin@liafa.jussieu.fr, straubin@cs.bc.edu,
denis@opus.cs.mcgill.ca

**Abstract**

We prove some results related to the generalized star-height problem. In this problem, as opposed to the restricted star-height problem, complementation is considered as a basic operator. We first show that the class of languages of star-height $\leq n$ is closed under certain operations (left and right quotients, inverse alphabetic morphisms, injective star-free substitutions). It is known that languages recognized by a commutative group are of star-height 1. We extend this result to nilpotent groups of class 2 and to the groups that divide a semidirect product of a commutative group by $(\mathbb{Z}/2\mathbb{Z})^n$. In the same direction, we show that one of the languages that was conjectured to be of star height 2 during the past ten years, is in fact of star height 1. Next we show that if a rational language $L$ is recognized by a monoid of the variety generated by wreath products of the form $M \circ (G \circ N)$, where $M$ and $N$ are aperiodic monoids, and $G$ is a commutative group, then $L$ is of star-height $\leq 1$. Finally we show that every rational language is the inverse image, under some morphism between free monoids, of a language of (restricted) star-height 1.

The determination of the star-height of a rational language is an old problem of formal language theory (see Brzozowski [1], for an historical survey). The restricted star-height problem has been recently solved by Hashiguchi [4], but here we are interested in that aspect of the problem concerning *generalized star-height*, in which complementation is considered as a basic operator. Thus, in the rest of this paper, the word "star-height" will always refer to generalized star-height.

The aim of this paper is to present some new results related to the star-height problem : "Is there an algorithm to compute the star-height of a given rational language" (this language can be given, for instance, by a rational expression). The star-height problem seems to be extremely difficult, and very little is known on the subject. For instance, it is not yet known whether there

*Research supported by the PRC Mathématiques et Informatique.
†Research supported by National Science Foundation Grant CCR-8700700.
‡Research supported by NSERC Grant A4546 and FCAR Grant 86-EQ-2933

is a language of star-height $> 1$. The most important known result is the theorem of Schützenberger [10] that gives an algebraic characterization of the languages of star-height 0 (also called star-free languages). A language is star-free if and only if its syntactic monoid is aperiodic (or group-free). This theorem has greatly influenced subsequent research. First, since group-free is equivalent to star-free, it is natural to search for candidates having star-height 2 or more among languages whose syntactic monoid is a group (or equivalently, that are accepted by a permutation automaton). The intuitive idea is that a "complex" group should recognize "complex" languages. But what kind of complexity is required for the group ? Henneman [5] was the first to study this problem in a systematic way. He showed that any language recognized by a commutative group is of star-height $\leq 1$, and gave some upper bounds on the star-heights of languages recognized by various classes of groups. In the quest for a language of star-height $> 1$, the next level of complexity was nilpotent groups of class 2, and such candidates were actually proposed in Brzozowski [1]. The combinatorial structure of the languages recognized by nilpotent groups is related to the generalized binomial coefficients introduced by Eilenberg [3], which count the number of times that a word u appears as a subword (in the sense of subsequence) of a word $v$. A precise description, involving the class of nilpotency of the group, is given in Thérien [14]. Let $L(u, k, n)$ denote the set of words $w$ in which the number of appearances of $u$ as a subword is congruent to $k$ mod $n$. Then a language is recognized by a nilpotent group of class $c$ if and only if it is a boolean combination of languages $L(u, k, n)$ where $|u| \leq c$. Thus, the problem of finding the star-height of languages recognized by nilpotent groups reduces to finding the star-height of the languages $L(u, k, n)$. Henneman's result mentioned above is a consequence of the fact that the star-height of $L(u, k, n)$ is 1 if $u$ is a word of length 1. Here we show that the star-height of $L(u, k, n)$ is at most 1 if $u$ is a word of length $\leq 2$; this corresponds to the case of nilpotent groups of class 2. We were not able to treat completely the case of words of length 3. However, we prove that the star-height of $L(u, k, n)$ is at most 1 if $u$ is a word of length $\leq 3$ and $n$ is a *square-free* integer. This covers in particular the case of $L(abc, 0, 2)$, which was proposed as a candidate for having star-height 2 in Brzozowski [1].

Using slightly different techniques, we give some other classes of monoids or groups that recognize only languages of star-height $\leq 1$. For instance the groups that divide a semidirect product of a commutative group by $(\mathbb{Z}/2\mathbb{Z})^n$, and the monoids that divide a wreath product of the form $M \circ (G \circ N)$, where $M$ and $N$ are aperiodic monoids, and $G$ is a commutative group. In particular, every language recognized by a group of order less than 12 is of star-height at most 1.

We also investigate the closure properties of languages of star height $\leq n$, for a given $n$. By definition, these classes are closed under boolean operations and concatenation. We show they are also closed under left and right quotients, star-free injective substitutions and inverse alphabetic morphisms. On the other hand, we prove that every rational language is the inverse image, under some morphism between free monoids, of a language of (restricted) star-height 1. In particular, if languages of star-height $\leq 1$ are closed under inverse morphisms, every rational language is of star-height $\leq 1$.

The paper is divided into eight sections. Section 1 contains a precise definition of the star-height problem. Section 2 gives some basic results about monoids and varieties and the known results on star-height are presented in Section 3. Operations preserving star-height are the subject of Section 4. In Section 5 we recall some basic facts about sequential functions and wreath products. Section 6 contains most of the technical results. The first lemma of this section, called the transfer lemma is of special interest. For instance, it gives an easy proof of the results of Thomas [15]. The main results of the paper are presented in Section 7. Section 8 concludes the paper with some further comments on the problem of finding a language of star-height 2.

The results of this paper have been announced in Pin *et al.* [9].

# 1 The generalized star-height problem.

Given a finite alphabet $A$, the (*extended*) *rational expressions* over $A$ are defined recursively as follows.

(a) $\emptyset$, 1 and $a$ (for every $a \in A$) are rational expressions.

(b) If $E$ and $F$ are rational expressions, so are $(E \cup F)$, $(EF)$, $E^c$ and $E^*$.

The (*generalized*) *star-height* $h(E)$ of a rational expression $E$ is defined recursively by

(a) $h(\emptyset) = 0$, $h(1) = 0$ and, for every $a \in A$, $h(a) = 0$.

(b) $h(E \cup F) = h(EF) = \max(h(E), h(F))$, $h(E^c) = h(E)$ and $h(E^*) = h(E) + 1$.

The *value* $v(E)$ of a rational expression $E$ is the language represented by $E$ ($E^c$ stands for the complement of $E$). More formally, $v$ is recursively defined by

(a) $v(\emptyset) = \emptyset$, $v(1) = \{1\}$ and, for every $a \in A$, $v(a) = \{a\}$.

(b) $v(E \cup F) = v(E) \cup v(F)$, $v(EF) = v(E)v(F)$, $v(E^c) = A^* \setminus v(E)$ and $v(E^*) = (v(E))^*$.

The (*generalized*) *star-height* $h(L)$ of a rational language $L$ is the minimum of the star-heights of all rational expressions representing $L$. One can give another description of the star-height, which follows directly from the definition.

**Proposition 1.1** *For every $n \geq 0$, the set of languages of star-height $\leq n+1$ is the smallest class of languages containing the languages of the form $L$ or $L^*$, where $h(L) \leq n$, and closed under boolean operations and concatenation product.*

The star-height problem is : Is there an algorithm to compute the star-height of a given rational language? (This language can be given, for instance, by a rational expression). The aim of this paper is to present some new results related to this problem. We introduce some rather deep techniques to establish that a language is of star-height 1. For instance, we show that if the syntactic monoid of a language is a nilpotent group of class two, then L is of star-height $\leq 1$.

Similar results hold for various varieties of finite monoids. Another consequence of our results is that one of the languages presented in Brzozowski [1] as a possible candidate for star-height 2 is in fact of star-height 1. This result may appear rather modest, but is in fact highly non-trivial. For instance, if one really wanted to write down the expression of star-height 1 obtained for this language, ten pages would probably not suffice! This means that there is probably no hope to find an expression of star-height 1 for this language by brute force.

## 2 Monoids and varieties.

Monoids often permit one to give an algebraic solution of problems about rational languages. Recall that a monoid $M$ recognizes a language $L$ of $A^*$ if there exists a monoid morphism $\eta : A^* \to M$ and a subset $P$ of $M$ such that $L = P\eta^{-1}$. The *syntactic congruence* of a language $L$ of $A^*$ is the congruence $\sim_L$ defined by

$$u \sim_L v \text{ if and only if, for every } x, y \in A^*, \ (xuy \in L \Leftrightarrow xvy \in L).$$

The quotient $M(L) = A^*/{\sim_L}$ is the *syntactic monoid* of $L$. In fact the syntactic monoid of $L$ is the "smallest" monoid that recognizes $L$. More precisely, $M(L)$ recognizes $L$, and $M(L)$ divides (that is, is a quotient of a submonoid of) every monoid that recognizes $L$. As is well-known, a language is rational if and only if it is recognized by a finite monoid (or equivalently, if and only if its syntactic monoid is finite).

Note that, given a rational expression $E$, one can effectively compute the syntactic monoid of the language $L$ represented by $E$. Indeed, there exist standard algorithms to compute the minimal automaton of $L$. Now, the syntactic monoid of $L$ is simply the transition monoid of this minimal automaton. The reader is referred to [3, 6, 8] for further details.

A *variety of monoids* is a class of finite monoids closed under taking submonoids, quotients (that is, morphic images) and finite direct products. For instance, the class of aperiodic monoids, considered in the next section, is a variety of monoids, denoted by $\mathbf{A}$. The class of finite groups is also a variety of monoids, denoted by $\mathbf{G}$. This variety contains some well-known subvarieties, for instance the variety of finite commutative groups, and, for every $n > 0$, the variety of nilpotent groups of class $n$. Recall that the lower central series of a group $G$ is defined as $G_1 = G$ and $G_{i+1} = [G_i, G]$, where $[H, K]$ denotes the subgroup of $G$ generated by all elements $h^{-1}k^{-1}hk$, $h \in H$, $k \in K$. A group $G$ is nilpotent of class $c$ if $G_c \neq \{1\}$ and $G_{c+1} = \{1\}$.

Let $M$ and $N$ be two monoids. We write $M$ additively (although $M$ is not assumed to be commutative) and $N$ multiplicatively. In particular, we denote by 0 and 1 the identities of $M$ and $N$ respectively. A (*left*) *action* of $N$ on $M$ is a function

$$\begin{aligned} N \times M &\to M \\ (n, m) &\to n \cdot m \end{aligned}$$

satisfying for every $m, m_1, m_2 \in M$ and $n, n_1, n_2 \in N$,

$$n \cdot (m_1 + m_2) = n \cdot m_1 + n \cdot m_2, \qquad n_1 \cdot (n_2 \cdot m) = (n_1 n_2) \cdot m,$$
$$n \cdot 0 = 0, \qquad 1 \cdot m = m$$

Given an action of $N$ on $M$, the *semidirect product* $M * N$ is the monoid defined on $M \times N$ by the multiplication

$$(m, n)(m', n') = (m + nm', nn').$$

The *wreath product* $M \circ N$ is the monoid defined on the set $M^N \times N$ by the multiplication given by the following formula (where $f_1$, $f_2$ are applications from $M$ into $N$, and $n_1$, $n_2$ are elements of $N$)

$$(f_1, n_1)(f_2, n_2) = (f, n_1 n_2)$$

where $f$ is the application from $M$ into $N$ defined, for all $n \in N$, by $nf = nf_1 + (nn_1)f_2$.

Given two varieties of monoids $\mathbf{V}$ and $\mathbf{W}$, we denote by $\mathbf{V} * \mathbf{W}$ the variety generated by all semidirect products of a monoid of $\mathbf{V}$ by a monoid of $\mathbf{W}$, which is also the variety generated by all wreath products of a monoid of $\mathbf{V}$ by a monoid of $\mathbf{W}$.

One of the goals of variety theory, introduced by Eilenberg, is to describe the class of languages whose syntactic monoids belong to a given variety of monoids. We shall briefly recall such a description for the varieties of commutative and nilpotent groups in this section, and for the variety $\mathbf{A}$ in the next section.

A word $u = a_1 a_2 \cdots a_n$ is a *subword* of a word $v$ if $v$ can be factored as $v = v_0 a_1 v_1 \cdots a_n v_n$. For instance, $ab$ is a subword of $cacbc$. Given two words $u$ and $v$, we denote by $\binom{v}{u}$ the number of times that $u$ appears as a subword of $v$.

More formally, if $u = a_1 a_2 \cdots a_n$, then

$$\binom{v}{u} = \mathrm{Card}\{(v_0, v_1, \ldots, v_n) \mid v_0 a_1 v_1 \cdots a_n v_n = v\}$$

Observe that if $u$ is a letter $a$, then $\binom{v}{a}$ is simply the number of occurrences of the letter $a$ in $v$, also denoted by $|v|_a$. More generally, if $B$ is a subset of the alphabet $A$, we put $|v|_B = \sum_{b \in B} |v|_b$. Finally, $|v| = |v|_A$ denotes the length of $v$.

For every word $u$ of $A^*$, and for any integers $k$ and $n$ such that $0 \leq k < n$, we put

$$L(u, k, n) = \left\{ v \in A^* \,\middle|\, \binom{v}{u} \equiv k \mod n \right\}.$$

We can now state

**Theorem 2.1** [3, 6, 8] *Let $L$ be a recognizable language. Then the following conditions are equivalent :*

(1) *$L$ is recognized by a finite commutative group,*

(2) *the syntactic monoid of $L$ is a finite commutative group,*

(3) *$L$ is a boolean combination of languages of the form $L(a, k, n)$, where $a$ is a letter, and $0 \leq k < n$.*

Note that one could take $0 < k < n$ in condition (3), since

$$L(a, 0, n) = A^* \setminus \bigcup_{0<k<n} L(a, k, n).$$

**Theorem 2.2** [14] *Let $L$ be a recognizable language. Then the following conditions are equivalent :*

(1) *$L$ is recognized by a finite nilpotent group of class $m$,*

(2) *the syntactic monoid of $L$ is a finite nilpotent group of class $m$,*

(3) *$L$ is a boolean combination of languages of the form $L(u, k, n)$, where $|u| \leq m$, and $0 \leq k < n$.*

# 3 Some known results on star-height.

The first major result on star-height was the algebraic characterization of the languages of star-height 0, also called star-free languages, obtained by Schützenberger in 1965.

Recall that a finite monoid $M$ is *aperiodic* if for every $x \in M$, there exists an integer $n$ such that $x^n = x^{n+1}$. Again this property can be effectively tested.

**Theorem 3.1** Schtzenberger, 1965) *Let $L$ be a rational language. Then the following conditions are equivalent :*

(1) *$L$ is recognized by a finite aperiodic monoid,*

(2) *$M(L)$ is aperiodic,*

(3) *$h(L) = 0$.*

**Corollary 3.2** *There is an algorithm to decide if a given rational language has star-height* 0.

The complexity of this algorithm is analyzed in Stern [11]. Given a finite deterministic automaton $\mathcal{A}$, deciding whether $\mathcal{A}$ recognize a star-free set can be solved in polynomial space. It is also the complement of an NP-hard problem.

There is a very elementary example for which a direct proof is possible.

**Lemma 3.3** *Let $B$ be a subset of the alphabet $A$. Then $h(B^*) = 0$.*

**Proof.** Indeed, we have $B^* = (\emptyset^c(A \setminus B)\emptyset^c)^c$ and thus $h(B^*) = 0$. $\square$

Theorem 3.1 shows that there exist some languages of star-height 1, for instance $(aa)^*$ (it is not difficult to verify that the syntactic monoid of this language is not aperiodic). Theorem 3.1 also suggests that one study the star-height problem through properties of the syntactic monoid. In this direction, Henneman [5] has studied the languages whose syntactic monoids are groups. The case of commutative groups is especially interesting.

**Theorem 3.4** Henneman [5] *A language recognized by a finite commutative group is of star-height* $\leq 1$.

**Proof.** By Theorem 2.1, $L$ is a boolean combination of languages of the form $L(a,k,n)$, where $a$ is a letter, and $0 \leq k < n$. But

$$L(a,k,n) = (B^*a)^k((B^*aB^*)^n)^* \text{ where } B = A \setminus \{a\}.$$

Now $h(B^*) = 0$ by Lemma 3.3. It follows that $h(L(a,k,n)) \leq 1$ and finally $h(L) \leq 1$ as required. □

**Corollary 3.5** *A language recognized by a finite commutative monoid is of star-height* $\leq 1$.

**Proof.** Indeed, by a result of Eilenberg [3], a language of $A^*$ recognized by a commutative monoid is a boolean combination of languages of the form $(B^*a)^kB^*$ (where $B = A \setminus \{a\}$ and $k \geq 0$) and of languages recognized by a commutative group. □

Non-trivial examples of languages of star-height 1 were given by Thomas [15]. Let $A = \{a,b\}$ and, for each $n \geq 0$, put $x^n = a^nb$. Then the set $X = \{x_n \mid n \geq 0\}$ is a prefix code such that $X^* = A^*b \cup \{1\}$. In particular, every word of $X^*$ admits a unique factorization as a product of words of $X$. Now, let $W(h,k,r,m)$ be the set of words $w$ of $X^*$ such that, in the factorization of $w$, the number of factors $x_n$ with $n \equiv r \mod m$ is congruent to $h \bmod k$. Then we have

**Theorem 3.6** (Thomas [15]) *For every $h$, $k$, $r$, $m$, the languages $W(h,k,r,m)$ are of star-height at most* 1.

# 4 Operations that preserve star-height.

By definition, the class of all languages of star-height $\leq n$ is closed under boolean operations and concatenation. In this section, we show that this class is also closed under other operations : left and right quotients, star-free injective substitutions, and inverse alphabetic morphisms.

If $K$ and $L$ are two languages of $A^*$, we call the left (or right) *quotient* or *residual* of $L$ by $K$ the language $K^{-1}L$ (or $LK^{-1}$) which is defined by

$$K^{-1}L = \{v \in A^* \mid \text{there exists } u \in K \text{ such that } uv \in L\}$$
$$LK^{-1} = \{v \in A^* \mid \text{there exists } u \in K \text{ such that } vu \in L\}$$

**Proposition 4.1** *For every rational language $L$ of $A^*$, and for every language $K$ of $A^*$, $h(K^{-1}L) \leq h(L)$ and $h(LK^{-1}) \leq h(L)$. In particular, for every $n \geq 0$, the set of languages of star height $\leq n$ is closed under left and right quotients.*

**Proof.** We only give the proof for left quotients, since the proof for right quotients is dual. First of all, it is well-known that for a rational language $L$, every quotient $K^{-1}L$ is a finite union of languages of the form $u^{-1}L$, where $u$ is a word. Furthermore, since $(uv)^{-1}L = v^{-1}(u^{-1}L)$, it suffices by induction to show that $h(a^{-1}L) \leq h(L)$ for every letter $a \in A$. We prove this result by induction on $n = h(L)$. This is true for $n = 0$, since by the theorem of Schützenberger, star-free languages form a variety of languages. Assume that the result holds for $n$, and let $\mathcal{C}$ be the class of all the languages $L$ of $A^*$ such that $h(L) \leq n+1$ and $h(a^{-1}L) \leq n+1$ for every letter $a \in A$. Let $K$ be a language such that $h(K) \leq n$. Then $K \in \mathcal{C}$ by induction, and also $K^* \in \mathcal{C}$. Indeed, $h(K^*) \leq n+1$, and, for every letter $a \in A$, $a^{-1}K^* = (a^{-1}K)K^*$ so that $h(a^{-1}K^*) \leq \max\{h(a^{-1}K), h(K^*)\} \leq n+1$. Thus $\mathcal{C}$ contains every language of the form $K$ or $K^*$ where $h(K) \leq n$. Finally assume that $K, K_1, K_2 \in \mathcal{C}$. Then $h(K), h(K_1), h(K_2) \leq n+1$ and hence $h(K^c), h(K_1 \cup K_2), h(K_1K_2) \leq n+1$. Furthermore $h(a^{-1}(K^c)) \leq n+1$ and $h(a^{-1}(K_1 \cup K_2)) \leq n+1$ since $a^{-1}(K_1 \cup K_2) = a^{-1}K_1 \cup a^{-1}K_2$ and $a^{-1}K^c = (a^{-1}K)^c$, so that $K_1 \cup K_2 \in \mathcal{C}$ and $K^c \in \mathcal{C}$. Similarly, $a^{-1}(K_1K_2) = (a^{-1}K_1)K_2$ if $1 \notin K$ and $a^{-1}(K_1K_2) = (a^{-1}K_1)K_2 \cup a^{-1}K_2$ if $1 \in K_1$ and thus $K_1K_2 \in \mathcal{C}$. Therefore, by Proposition 1.1, $\mathcal{C}$ contains all the languages of star height $\leq n+1$, and this concludes the proof. $\square$

A *substitution* $\sigma : A^* \to B^*$ is a relation on $A^* \times B^*$ which induces a map from $A^*$ into $\mathcal{P}(B^*)$ such that $1\sigma = \{1\}$ and, for every $u, v \in A^*$, $(uv)\sigma = (u\sigma)(v\sigma)$. A substitution is *rational* if, for every $a \in A$, as is a rational language. This implies in particular that, for every rational language $L$, the language

$$L\sigma = \bigcup_{u \in L} u\sigma$$

is a rational language. All the substitutions considered in this article will be rational substitutions.

A substitution $\sigma : A^* \to B^*$ is *injective* if for every $u, v \in A^*$, $u\sigma \cap v\sigma \neq \emptyset$ implies $u = v$. (Note that this definition is compatible with the definition of an injective relation, but does not mean that $\sigma$ induces an injective function from $A^*$ into $\mathcal{P}(B^*)$). The next proposition provides useful examples of injective substitutions. Recall that a subset $X$ of $A^+$ is a *code* if, for every $x_1, \ldots, x_n, y_1, \ldots, y_m \in X$, $x_1x_2 \ldots x_n = y_1 \ldots y_m$ implies $n = m$ and $x_i = y_i$ for $i = 1, \ldots, n$.

**Proposition 4.2** *Let $\sigma : A^* \to B^*$ be a substitution. Then if the sets $a\sigma$ (for $a \in A$) are pairwise disjoint, and if $A\sigma$ is a code, then $\sigma$ is injective.*

**Proof.** . Assume that $x \in (a_1 \cdots a_r)\sigma \cap (b_1 \cdots b_s)\sigma$ for some $a_1, \ldots, a_r, b_1, \ldots, b_s \in A$. Then there exist $x_1 \in a_1\sigma, \ldots, x_r \in a_r\sigma$, $y_1 \in b_1\sigma, \ldots, y_s\sigma \in b_s\sigma$

such that $x = x_1 \cdots x_r = y_1 \cdots y_s$. Since $A\sigma$ is a code, it follows that $r = s$ and $x_1 = y_1, \ldots, x_s = y_s$. Thus $x_1 \in a_1\sigma \cap b_1\sigma, \ldots, x_s \in a_s\sigma \cap b_s\sigma$ and hence $a_1 = b_1, \ldots, a_s = b_s$ since the sets $a\sigma$ are pairwise disjoint. Thus $\sigma$ is injective. $\square$

The converse of Proposition 4.2 is false. For instance let $\sigma : a^* \to \{a, b\}^*$ be the substitution defined by $a\sigma = \{a, ab, ba\}$. Then $\sigma$ is injective since $x \in a^n\sigma$ implies $n = |x|_a$, but $a\sigma$ is not a code since $a(ba) = (ab)a$. The following proposition summarizes some well-known properties of substitutions.

**Proposition 4.3** *Let $\sigma : A^* \to B^*$ be a substitution. Then for every language $L, L_1, L_2 \subset A^*$, we have*

(1) $(L_1 \cup L_2)\sigma = L_1\sigma \cup L_2\sigma$

(2) $(L_1L_2)\sigma = (L_1\sigma)(L_2\sigma)$

(3) $L^*\sigma = (L\sigma)^*$

(4) *If furthermore, $\sigma$ is injective, then $(L_1 \cap L_2)\sigma = L_1\sigma \cap L_2\sigma$ and $(L_1 \setminus L_2)\sigma = L_1\sigma \setminus L_2\sigma$.*

**Proof.** (1), (2) and (3) are obvious. Assume that $\sigma$ is injective. Clearly $(L_1 \cap L_2)\sigma \subset L_1\sigma \cap L_2\sigma$. Conversely, let $v \in L_1\sigma \cap L_2\sigma$. Then $v \in u_1\sigma \cap u_2\sigma$ for some $u_1 \in L_1$ and $u_2 \in L_2$. Since $\sigma$ is injective, $u_1 = u_2 \in L_1 \cap L_2$ and thus $v \in (L_1 \cap L_2)\sigma$. This proves that $(L_1 \cap L_2)\sigma = L_1\sigma \cap L_2\sigma$.

In particular, $(L_1 \setminus L_2)\sigma \cap L_2\sigma = \emptyset\sigma = \emptyset$. It follows that $(L_1 \setminus L_2)\sigma \subset L_1\sigma \setminus L_2\sigma$. But since $L_1\sigma = (L_1 \setminus L_2)\sigma \cup L_2\sigma$, we have $(L_1 \setminus L_2)\sigma = L_1\sigma \setminus L_2\sigma$. $\square$

A substitution $\sigma$ is *star-free* if for every star-free language $L$, $L\sigma$ is also star-free. For instance, $\sigma : \{a, b\}^* \to \{a, b, c\}^*$ defined by $a\sigma = \{a, ab\}$, $b\sigma = \{c, bc\}$ is an injective star-free substitution. It is decidable whether or not an effectively given rational injective substitution is star-free.

**Proposition 4.4** *Let $\sigma : A^* \to B^*$ be an injective substitution. Then $\sigma$ is star-free if and only if it satisfies the following conditions :*

(1) *for every $a \in A$, $a\sigma$ is star-free,*

(2) *$(A\sigma)^*$ is star-free.*

**Proof.** Since $(A\sigma)^* = A^*\sigma$, the condition is necessary. Conversely, let $\sigma$ be an injective substitution that satisfies (1) and (2). Let $\mathcal{S}$ be the set of all rational languages $L$ of $A^*$ such that $L\sigma$ is star-free. Since $1\sigma = 1$, $\{1\} \in \mathcal{S}$, and by (1), $\mathcal{S}$ contains every letter. Furthermore, Proposition 4.2 shows that $\mathcal{S}$ is closed under union, difference and concatenation. Finally if $L \in \mathcal{S}$, then $(A^* \setminus L)\sigma = (A\sigma)^* \setminus L\sigma$ is star-free by (2) and thus $A^* \setminus L \in \mathcal{S}$. Thus $\mathcal{S}$ is also closed under complementation and hence contains all star-free languages. $\square$

There exist injective substitutions which satisfies (2) but not (1). For instance, the substitution $\sigma : \{a,b\}^* \to \{a,b\}^*$ defined by $a\sigma = b(a^2)^*$ and $b\sigma = b(a^2)^*a$. But if $\sigma$ is a morphism, condition (1) is always satisfied since $a\sigma$ is a single word in this case. Furthermore, since $\sigma$ is injective as a substitution, $A\sigma$ is a finite code. It is known [3] that $(A\sigma)^*$ is star-free if and only if $A\sigma$ is a *pure code* (a code $X$ is pure if $u^n \in X^*$ for some $n > 0$ implies $u \in X^*$). An injective morphism $\varphi$ such that $A\varphi$ is a pure code is called a *pure coding*. Therefore

**Corollary 4.5** *An injective morphim is star-free if and only if it is a pure coding.*

In fact, as one of the referees pointed out, one can show that a star-free morphism which does not map every word on the empty word is injective. We shall not use this stronger result in this paper. We can now state the main result of this section.

**Theorem 4.6** *Let $\sigma : A^* \to B^*$ be a star-free injective substitution. Then for every rational language $L$, $h(L\sigma) \le h(L)$. In particular, for every $n \ge 0$, the set of languages of star-height $\le n$ is closed under star-free injective substitutions.*

**Proof.** Let $\sigma$ be a star-free substitution. For every $a \in A$, let $E_a$ be a star-free expression representing $a\sigma$ and let $E_*$ be a star-free expression for $(A\sigma)^*$. We first extend $\sigma$ to rational expressions as follows:

$$\begin{aligned}
\emptyset\sigma = \emptyset, 1\sigma = 1 \text{ and } a\sigma &= E_a \text{ for every } a \in A \\
(E_1 \cup E_2)\sigma &= E_1\sigma \cup E_2\sigma \\
(E_1E_2)\sigma &= (E_1\sigma)(E_2\sigma) \\
(E^c)\sigma &= (E_* \setminus E\sigma) \\
E^*\sigma &= (E\sigma)^*
\end{aligned}$$

It is not difficult to prove by induction on $E$ that

(a) $v(E\sigma) = (v(E))\sigma$

(b) $h(E\sigma) \le h(E)$.

Now assume that $h(L) \le n$. Then there exists an expression $E$ such that $h(E) \le n$ and $v(E) = L$. By (a), $E\sigma$ is an expression representing $L\sigma$ and by (b), $h(E\sigma) \le n$. Thus $h(L\sigma) \le n$. □

Recall that a morphism $\varphi : A^* \to B^*$ is *alphabetic* if, for every letter $a \in A$, $a\varphi$ is either a letter of $B$ or the empty word. Then we can state

**Corollary 4.7** *For every $n \ge 0$, the set of languages of star-height $\le n$ is closed under inverse alphabetic morphisms.*

**Proof.** Let $\varphi : A^* \to B^*$ be an alphabetic morphism. Define a relation $\sigma : B^* \to A^*$ by setting $1\sigma = 1$, and for every word $u \in B^+$, $u\sigma = u\varphi^{-1}$. We claim that $\sigma$ is an injective star-free substitution. First of all, since $\varphi$ is alphabetic, $\sigma$ is an injective substitution. Put $C = \{a \in A^* \mid a\varphi = 1\}$, and, for every letter $b \in B$, $C_b = \{a \in A \mid a\varphi = b\}$. Then for every $b \in B$, $b\varphi^{-1} = C^* C_b C^*$ is star-free by Lemma 3.3. Finally, $B^*\sigma = A^*$ is star-free and thus $\sigma$ is star-free by Proposition 4.4. Thus, by Theorem 4.6, if $L$ is a language of $A^*$ such that $h(L) \leq n$, then $h(L\sigma) \leq n$. Now if 1 is not in $L$, then $L\varphi^{-1} = L\sigma$ and hence $h(L\varphi^{-1}) \leq n$. If $1 \in L$, we have $L\varphi^{-1} = C^* \cup (L \setminus \{1\})\sigma$ and since $C^*$ is star-free, $h(L\varphi^{-1}) \leq h((L \setminus \{1\})\sigma)$. Now by Theorem 4.6, $h((L \setminus \{1\})\sigma) \leq h(L\ \{1\}) = h(L)$. Therefore $h(L\varphi^{-1}) \leq n$ as required. $\square$

# 5 Sequential functions and wreath product

In this section we review the definition of sequential functions and their relations with wreath products.

Recall that a (*left sequential*) *transducer* $\mathcal{T} = (Q, A, B, \cdot, *, q_0)$ consists of an input alphabet $A$, an output alphabet $B$, a finite set of states $Q$, an initial state $q_0 \in Q$, and two functions

$$\begin{array}{ccc} Q \times A \to Q & \qquad & Q \times A \to B \\ (q, a) \to q \cdot a & & (q, a) \to q * a \end{array}$$

called the next state function and the output function, respectively. These functions are extended to $Q \times A^*$ by setting, for $u \in A^*$ and $a \in A$,

$$\begin{array}{ccc} q \cdot 1 = q & \qquad & q \cdot (ua) = (q \cdot u) \cdot a \\ q * 1 = 1 & & q * (ua) = (q * u)((q \cdot u) * a) \end{array}$$

The (partial) function $\sigma : A^* \to B^*$ realized by $\mathcal{T}$ is defined by

$$u\sigma = q_0 * u.$$

A *sequential function* is a function realized by such a transducer.

We shall use the classical "state transition diagram" to represent the deterministic automaton $\mathcal{A} = (Q, A, \cdot, q_0)$. For instance, the fact that $q \cdot a = q'$ is represented by an edge

$$q \xrightarrow{a} q'.$$

Assume that $\sigma$ is a total function, or, equivalently, that $\mathcal{A}$ is a complete automaton (this will be the case in all the examples considered in this paper). Then every word $u = a_1 \cdots a_k$ defines a unique path

$$p(u) = (q_0, a_1, q_1)(q_1, a_1, q_2) \cdots (q_{k-1}, a_k, q_k)$$

where $q_0$ is the initial state and $q_{i+1} = q_i \cdot a_{i+1}$ for $0 \leq i \leq k-1$. Therefore we may use without ambiguity the shorter notation

$$p(u) = (q_0, a_1)(q_1, a_2) \cdots (q_k - 1, a_k) \tag{1}$$

Now we have by definition

$$u\sigma = (q_0 * a_1)(q_1 * a_2) \cdots (q_{k-1} * a_k) \tag{2}$$

and it follows from (1) and (2) that

$$|u\sigma|_b = \sum_{q*a=b} |p(u)|_{(q,a)}$$

Thus counting the number of occurrences of a given letter in $u\sigma$ reduces to counting the number of occurrences of a given edge in $p(u)$. Therefore, it suffices to count modulo $n$ the number of occurrences of a given edge in the path defined by a word $u$. We shall discuss this problem in detail in Section 6.

Sequential functions are intimately related to wreath products. Indeed let $M(\sigma)$ be the transition monoid of the automaton $\mathcal{A}$ defined above. Then if a language $L \subset B^*$ is recognized by a monoid $M$, $L\sigma^{-1}$ is recognized by the wreath product $M \circ M(\sigma)$. There is a partial converse to this result. Let $N \circ M$ be a wreath product and let $\eta : A^* \to N \circ M$ be a morphism recognizing a language $L$. We denote by $\pi : N \circ M \to M$ the natural projection and we put $\varphi = \eta\pi$ and $B = M \times A$. Then we have

**Proposition 5.1** (Wreath product principle [12]) *If $L$ is recognized by $\eta : A^* \to N \circ M$, then $L$ is a boolean combination of languages of the form $X \cap Y\sigma^{-1}$ where $X \subset A^*$ is recognized by $M$, $Y \subset B^*$ is recognized by $N$ and $\sigma : A^* \to B^*$ is the sequential function defined by*

$$(a_1 \cdots a_r)\sigma = (1, a_1)(a_1\varphi, a_2) \cdots ((a_1 \cdots a_{r-1})\varphi, a_r).$$

Note that the sequential function $\sigma$ is realized by the transducer $\mathcal{T} = (M, A, M \times A, \cdot, *, 1)$ where the next state function and the output function are defined, for every $m \in M$ and every $a \in A$, by $m \cdot a = m(a\varphi)$ and $m * a = (m, a)$.

We conclude this section by a definition. Given two varieties of monoids $\mathbf{V}$ and $\mathbf{W}$, we denote by $\mathbf{V} * \mathbf{W}$ the smallest variety of monoids containing all the wreath products of the form $M \circ N$ where $M \in \mathbf{V}$ and $N \in \mathbf{W}$. One can show that the operation $(\mathbf{V}, \mathbf{W}) \to \mathbf{V} * \mathbf{W}$ is an associative (but non-commutative!) operation on varieties.

# 6 Some languages of star-height 1

We have collected in this section the technical results that lead to the main results of the next section. The first of these results is called the Transfer Lemma because it is based on an identity in which the expression $b^*$ occurs on the left side but not on the right side, while the expression $a^*$ occurs on the right side but not on the left side. Thus, informally, stars have been transfered from $b^*$ to $a^*$.

**Lemma 6.1** (Transfer Lemma). *Let $L_0$ and $L_1$ be star-free languages of $A^*$ such that $L_0^*$ is star-free. Assume that the substitution $\sigma$ defined by $a\sigma = L_0$ and $b\sigma = L_1$ is injective and let $L = [(L_1^*L_0)^n]^*L_1^* = L_1^*[(L_0L_1^*)^n]^* = [(L_1^*L_0L_1^*)^n]^*$. Then $h(L) \leq 1$.*

**Proof.** Let $A = \{a, b\}$ be a two-letter alphabet. Since $|u|_a + |u|_b = |u|$, we have $|u|_a \equiv 0 \mod n$ if and only if there exists $r$ such that $0 \leq r < n$ and $|u| \equiv |u|_b \equiv r \mod n$. Therefore

$$L(a, 0, n) = \bigcup_{0 \leq r \leq n-1} [(A^n)^*A^r \cap L(b, r, n)]$$

Now, $L(a, 0, n) = (b \cup (ab^*)^{n-1}a)^* = ((b^*a)^n)^*b^*$ and $L(b, n, r) = (a^*b)^r(a \cup (ba^*)^{n-1}b)^*$ and thus we obtain the following formula:

$$((b^*a)^n)^*b^* = \bigcup_{0 \leq r \leq n-1} [(A^n)^*A^r \cap (a^*b)^r(a \cup (ba^*)^{n-1}b)^*] \tag{1}$$

By (1) and Proposition 4.2, we have

$$L = \bigcup_{0 \leq r \leq n-1} [((L_0 \cup L_1)^n)^*(L_0 \cup L_1)^r \cap (L_0^*L_1)^r(L_0 \cup (L_1L_0^*)^{n-1}L_1)^*]$$

Now $h(L_0^*) = 0$, $h(L_0) = 0$, $h(L_1) = 0$ and thus the above formula shows that $h(L) \leq 1$. □

The transfer Lemma is very useful since it permits one to remove a star level in an expression. As an example, we give a simple proof of Theorem 3.6. Put

$$L_0 = \{a^rb\} \quad \text{and} \quad L_1 = \{a^m, b, ab, \ldots, a^{r-1}b, a^{r+1}b, \ldots, a^{m-1}b\}.$$

Then one can verify that

$$W(h, k, r, m) = [(L_1^*L_0)^k]^*L_1^*(L_0L_1^*)^h \cap (A^*b \cup \{1\}).$$

Now $L_0$, $L_1$ and $L_0^*$ are star-free, and since $L_0 \cup L_1$ is a prefix code, the substitution $\sigma$ defined by $a\sigma = L_0$ and $b\sigma = L_1$ is injective by Proposition 4.2. Thus, by the Transfer Lemma, $[(L_1^*L_0)^k]^*L_1^*$ is of star-height 1, and so is $W(h, k, r, m)$.

In the sequel we encounter the following type of problem. Given a total function $\gamma : A^* \to \mathbb{N}$ and two integers $k, n$ such that $0 \leq k < n$, find the star height of the language

$$L(\gamma, k, n) = \{u \in A^* \mid u\gamma \equiv k \mod n\}.$$

Here is a first result to handle this problem.

**Proposition 6.2** *Let $c, c_1, \ldots, c_r \in \mathbb{Z} \setminus \{0\}$ and let $\gamma, \gamma_1, \ldots, \gamma_r : A^* \to \mathbb{N}$ be total functions such that, for every $u \in A^*$, $c(u\gamma) = c_1(u\gamma_1) + \ldots + c_r(u\gamma_r)$. Then for every $k, n$ such that $0 \leq k < n$, $L(\gamma, k, n)$ is a boolean combination of languages of the form $L(\gamma_i, k_i, c_n)$, where $0 \leq k_i < n$ and $1 \leq i \leq r$.*

**Proof.** First, $u\gamma \equiv k \mod n$ is equivalent to $c(u\gamma) \equiv ck \mod cn$. Now by definition, $c(u\gamma) = c_1(u\gamma_1) + \ldots + c_r(u\gamma_r)$ for every word $u \in A^*$. Therefore $c(u\gamma) \equiv ck \mod cn$ if and only if there exist $k_1, k_2, \ldots, k_r$ such that

(1) $u\gamma_i \equiv k_i \mod cn$ for $1 \leq i \leq r$, and

(2) $\sum_{1\leq i\leq r} c_i k_i = ck \mod cn$.

It follows that

$$L(\gamma, k, n) = \bigcup_{c_1k_1+\ldots+c_rk_r=c_k} \Big( \bigcap_{1\leq i\leq r} L(\gamma_i, k_i, cn) \Big). \qquad \square$$

We now come to the analysis of the situation already encountered in Section 5. Let $\mathcal{A} = (Q, A, \cdot, q_0)$ be a complete deterministic automaton (in which the set of final states is not specified). We have seen that every word $u = a_1 \cdots a_k$ defines a unique path $p(u) = (q_0, a_1)(q_1, a_2)\cdots(q_{k-1}, a_k)$. Let $q \in Q$, $a \in A$ and $0 \leq k < n$ be two integers. We would like to compute the star height of the language

$$L(\mathcal{A}, (q, a), k, n) = \{u \in A^* \mid |p(u)|_{(q,a)} \equiv k \mod n\}.$$

We start with two general results. Recall that an automaton $\mathcal{A}$ is transitive (or strongly connected) if for every $q_1, q_2 \in Q$, there exists a word $u$ such that $q_1 \cdot u = q_2$.

**Proposition 6.3** *Let $\mathcal{A}$ be a transitive deterministic automaton. Then the following equality holds for every $q \in Q$ and for all the integers $k, n$ such that $0 \leq k < n$: $h(L(\mathcal{A}, (q, a), k, n)) = h(L(\mathcal{A}, (q, a), 0, n))$.*

**Proof.** Let $r, s, t$ be words of minimal length such that $q_0 \cdot r = q$, $q \cdot as = q$ and $q \cdot t = q_0$. Put $w = r(as)^k t$. Then $q_0 \cdot w = q_0$ and $p(w)$ contains exactly $k$ occurrences of the edge $(q, a)$. We claim that

$$w^{-1}L(\mathcal{A}, (q, a), k, n) = L(\mathcal{A}, (q, a), 0, n).$$

Indeed, let $u \in L(\mathcal{A}, (q, a), k, n)$. Then since $q_0 \cdot w = q_0$, we have

$$|p(wu)|_{(q,a)} = |p(w)|_{(q,a)} + |p(u)|_{(q,a)} = k + |p(u)|_{(q,a)}.$$

Therefore $|p(u)|_{(q,a)} \equiv 0 \mod n$ if and only if $|p(wu)|_{(q,a)} \equiv k \mod n$, proving the claim. It follows by Proposition 4.1 that

$$h(L(\mathcal{A}, (q, a), k, n)) \leq h(\mathcal{A}, (q, a), 0, n))$$

and a dual argument proves the opposite inequality.

**Proposition 6.4** *Let $\mathcal{A} = (Q, A, ., q_0)$ be a transitive deterministic automaton. Assume that for every $u \in A^*$, $q \cdot u = q$ for some $q \in Q$ implies $q \cdot u = q$ for every $q \in Q$. Then for every $q \in Q$ and every $a \in A$, $h(L(\mathcal{A}, (q, a), 0, n)) = h(L(\mathcal{A}, (q_0, a), 0, n))$.*

**Proof.** Let $q \in Q$ and $a \in A$. Since $\mathcal{A}$ is transitive, there exists a word $v \in A^*$ such that $q_0 \cdot v = q$. We claim that

$$|p(vu)|_{(q,a)} = |p(v)|_{(q,a)} + |p(u)|_{(q_0,a)}.$$

Indeed, let $u = u_1 a u_2$ be a factorization of $u$ such that $q_0 \cdot u_1 = q_0$. Then $q \cdot u_1 = q$ by the hypothesis and $q_0 \cdot v u_1 = q \cdot u_1 = q$. Conversely, if $vu = (vu_1) a u_2$ with $q_0 \cdot v u_1 = q$, then $q \cdot u_1 = q$. Thus the claim holds and it follows that

$$L(\mathcal{A}, (q_0, a), 0, n) = v^{-1} L(\mathcal{A}, (q, a), k, n) \text{ where } k \equiv |p(v)|_{(q,a)} \mod n.$$

Therefore, by Propositions 4.1 and 6.3,

$$h(L(\mathcal{A}, (q_0, a), 0, n)) \leq h(L(\mathcal{A}, (q, a), 0, n))$$

and a dual argument would show the opposite inequality.

**Corollary 6.5** *Let $\mathcal{A} = (Q, A, \cdot, q_0)$ be a transitive deterministic automaton. If the transition monoid of $\mathcal{A}$ is commutative, then*

$$h(L(\mathcal{A}, (q, a), 0, n)) = h(L(\mathcal{A}, (q_0, a), 0, n))$$

*for every $q \in Q$ and every $a \in A$.*

We are now ready to treat our first example.

**Proposition 6.6** *Let $Q = \{0, 1, \ldots, n-1\}$ and let $\rho$ be the permutation on $Q$ defined by $q \cdot \rho = q + 1 \mod n$. Let $\mathcal{A} = (Q, A, \cdot, 0)$ be a complete deterministic automaton in which the action of each letter induces either the identity or the permutation $\rho$ on $Q$. Then for each $q \in Q$, for each letter $a \in A$ inducing the identity on $Q$, and for all the integers $k$ and $m$ such that $0 \leq k < m$, $h(L(\mathcal{A}, (q, a), k, m)) \leq 1$.*

**Proof.** The result is trivial if all the letters of $\mathcal{A}$ induce the identity on $Q$. Otherwise $\mathcal{A}$ is transitive and we may suppose $k = 0$ by Proposition 6.3 and $q = 0$ by Proposition 6.4. Set $B = \{a \in A \mid a \text{ induces the identity on } Q\}$ and let $C = A \setminus B$. Thus every letter of $C$ induces the permutation $\rho$ on $Q$ and $a \in B$. Put $D = B \setminus \{a\}$. The situation is summarized in Figure 1.

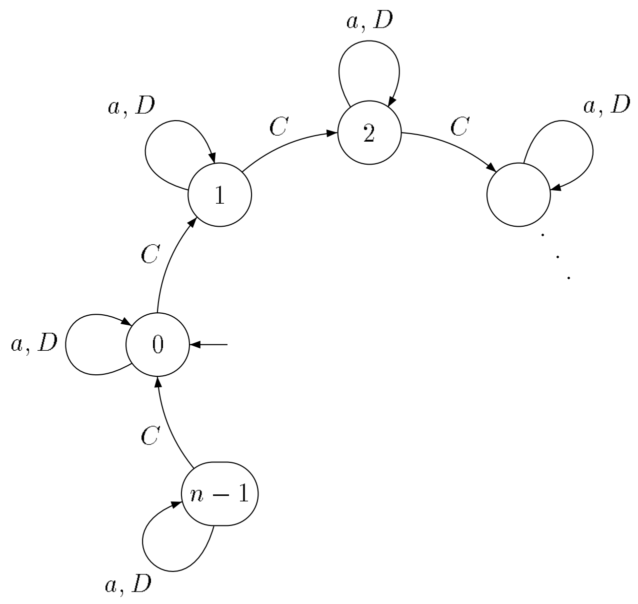



Figure 1:

We claim that

$$L(\mathcal{A}, (0,a), 0, m) = P^*[(aP^*)^m]^* S^{-1}$$

where

$$P = D \cup (CB^*)^{n-1}C \text{ and } S = \bigcup_{0 \le k < n} (CB^*)^k.$$

Indeed, let $u \in L(\mathcal{A}, (0,a), 0, m)$ and let $q = 0 \cdot u$. Since $\mathcal{A}$ is transitive, there exists at least one letter $c \in C$. Put $v = c^{n-q}$ if $q \neq 0$ and $v = 1$ if $n = 0$, so that $0 \cdot uv = q \cdot c^{n-q} = 0$. Then $v \in S$ and $uv \in P^*[(aP^*)^m]^*$ since $P$ is the set of all the words $x$ such that $p(x)$ is a simple loop from 0 to 0 containing no occurrence of $(0,a)$. Thus $u \in P^*[(aP^*)^m]^* S^{-1}$. Conversely, if $u \in P^*[(aP^*)^m]^* S^{-1}$, then $uv \in P^*[(aP^*)^m]^*$ for some $v \in S$. Put $q = 0 \cdot u$. Then $q \cdot v = 0$ and since $v \in S$, the path from $q$ to 0 defined by $v$ contains no occurrence of $(0,a)$. It follows that $|p(uv)|_{(0,a)} = |p(u)|_{(0,a)} \equiv 0 \mod m$. Thus $u \in L(\mathcal{A}, (0,a), 0, m)$, proving the claim.

Now $\{a\} \cup P$ is a prefix code and hence the substitution $\sigma : \{a,b\}^* \to A^*$ defined by $a\sigma = a$ and $b\sigma = P$ is injective by Proposition 4.2. Furthermore $a^*$ and $P$ are star-free by Lemma 3.3 and thus the Transfer Lemma can be applied to show that $h(P^*[(aP^*)^m]^*) \le 1$. It follows, by Proposition 4.1, that $h(L(\mathcal{A}, (0,a), 0, m)) \le 1$. □

A similar result holds for another type of automaton.

**Proposition 6.7** *If the transition monoid of a deterministic automaton* $\mathcal{A} = (Q, A, \cdot, 0)$ *is aperiodic, then* $h(L(\mathcal{A}, (q, a), k, n)) \le 1$ *for every* $q \in Q$, $a \in A$ *and for all the integers* $k$ *and* $n$ *such that* $0 \le k < n$.

**Proof.** Put $q' = q{\cdot}a$ and let $\mathcal{B}$ be the automaton deduced from $\mathcal{A}$ by erasing the transition $q{\cdot}a$ (so that $q{\cdot}a$ is undefined in $\mathcal{B}$). The transition monoid of $\mathcal{B}$ is also aperiodic. Otherwise, there exists a word $u$ inducing in $\mathcal{B}$ a non trivial permutation on a subset $K$ of $Q$. Therefore $u$ also induces in $\mathcal{A}$ a non trivial permutation on $K$, so that $\mathcal{A}$ is not aperiodic, a contradiction. For each $q_1, q_2 \in Q$, we define

$$K(q_1, q_2) = \{u \in A^* \mid q_1{\cdot}u = q_2 \in \mathcal{B}\}.$$

All these languages are star-free by Theorem 3.1. Now a simple inspection of the occurrences of $(q, a)$ in the path defined by a word in $\mathcal{A}$ leads to the following formulas, where $S = \bigcup_{s \in Q} K(q', s)$:

$$L(\mathcal{A}, (q, a), k, n) = \begin{cases} K(q_0, q)[(aK(q', q))^n]^*(aK(q', q))^{k-1}aS & \text{if } k > 0, \\ K(q_0, q)[(aK(q', q))^n]^*(aK(q', q))^{n-1}aS \cup \left(\bigcup_{s \in Q} K(q_0, s)\right) & \text{if } k = 0. \end{cases}$$

It follows that $h(L(\mathcal{A}, (q, a), k, n)) \le 1$. $\square$

We conclude this section by two slightly more technical results. Let $p$ be a prime number and let $\mathcal{A} = (Q, A, \cdot, q_0)$ be an automaton with $Q = (\mathbb{Z}/p\mathbb{Z})^r$, $q_0 = (0, \dots, 0)$, such that for every $a \in A$, there exists an $r$-tuple $v_a \in Q$ with $q{\cdot}a = q + v_a$ for all $q \in Q$. For instance, the automaton represented in Figure 2 is of this form

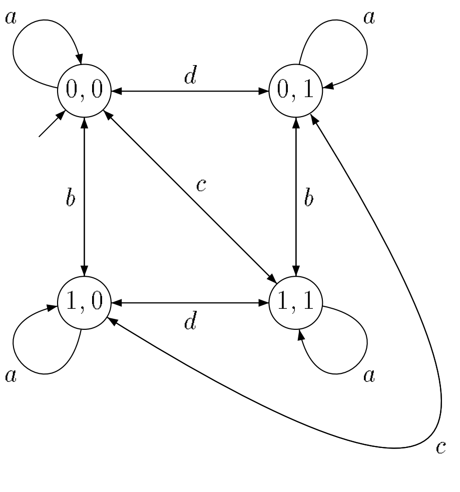



Figure 2:

We can now state

**Proposition 6.8** *Let $\mathcal{A} = (Q, A, \cdot, q_0)$ be one of the automata described above, and let $a$ be a letter of $A$ that induces the identity on $Q$. Then for all the integers $k$, $n$ such that $0 \le k < n$, $h(L(\mathcal{A}, (q, a), k, n)) \le 1$.*

**Proof.** If $\mathcal{A}$ is not transitive, we define a new automaton $\mathcal{B} = (Q, A \cup B, \cdot, q_0)$ as follows: $B = \{b_1, \ldots, b_r\}$ is a set of new letters and, for $1 \le i \le r$, $(q_1, \ldots, q_r) \cdot b_i = (q_1, \ldots, q_{i-1}, q_i + 1, q_{i+1}, \ldots, q_r)$. Now $\mathcal{B}$ is clearly transitive and we have

$$L(\mathcal{A}, (q, a), k, n)) = (L(\mathcal{B}, (q, a), k, n) \cap A^*)\varphi^{-1}$$

where $\varphi : A^* \to (A \cup B)^*$ is the natural morphism defined by $u\varphi = u$ for every $u \in A^*$. Now $A^*$ is a star-free subset of $(A \cup B)^*$ by Lemma 3.3 and $\varphi$ is an alphabetic morphism. Therefore, by Corollary 4.7, it suffices to show that $h(L(\mathcal{B}, (q, a), k, n)) \le 1$; that is, it suffices to prove the proposition for *transitive* automata. Therefore, we may suppose $k = 0$ by Proposition 6.2 and $q = q_0$ by Proposition 6.3.

We first treat the case $r = 1$. In this case every letter of $A$ induces a power of the cyclic permutation $\rho = (0, 1, \ldots, p-1)$. Let $P$ be the set of all the words $u$ such that $p(u)$ is a simple loop from 0 to 0 containing no occurrence of $(0, a)$. More formally

$$P = \{u \in A^* \mid 0 \cdot u = 0, |p(u)|_{(0,a)} = 0 \text{ and } \sum_{b \ne a} |p(u)|_{(0,b)} = 1\}$$

We claim that $P$ is star-free. Indeed $P$ is recognized by the automaton $\mathcal{C} = (\{0, 1, \ldots, p-1\} \cup \{f\}, A, \cdot, 0, \{f\})$, where $f$ is a new state and the transitions are the same as in $\mathcal{A}$ except for the transitions of the form $q \cdot a = 0$ which are replaced in $\mathcal{C}$ by $q \cdot a = f$.

Let $u \in A^+$ and let $q_1, \ldots, q_s$ be a sequence of states such that, in $\mathcal{C}$,

$$(1) \quad q_1 \cdot u = q_2, q_2 \cdot u = q_3, \ldots, q_{s-1} \cdot u = q_s, q_s \cdot u = q_1. \qquad (2)$$

The definition of the transitions of $\mathcal{C}$ implies that $q_1, \ldots, q_s \in \{1, \ldots, p-1\}$ and that relations (1) also hold in $\mathcal{A}$. But since $u$ induces a power of $\rho$ in $\mathcal{A}$, $s$ divides $p$. But $s < p$ and $p$ is prime, so that $s = 1$. Therefore the transition semigroup of $\mathcal{C}$ is aperiodic and $P$ is star-free by Theorem 3.1.

Since $\{a\} \cup P$ is a prefix code, Proposition 4.2 shows that the substitution $\sigma : \{a, b\}^* \to A^*$ defined by $a\sigma = a$ and $b\sigma = P$ is injective. Furthermore $a^*$ and $P$ are star-free and thus $h([(P^*aP^*)^n]^*) \le 1$ by the Transfer Lemma. Set, for each $q \in Q$,

$$S_q = \{u \in A^* \mid q \cdot u = 0 \text{ and } q \cdot v \ne 0 \text{ for any proper left factor } v \text{ of } u\}.$$

In particular $S_0 = \{1\}$. We claim that

$$L(\mathcal{A}, (0, a), 0, n) = [(P^*aP^*)^n]^* S^{-1} \text{ where } S = \bigcup_{q \in Q} S_q.$$

Indeed let $w \in L(\mathcal{A},(0,a),0,n)$ and let $q = 0{\cdot}w$. Since $\mathcal{A}$ is transitive, there exists a word $u \in A^*$ of minimal length such that $q{\cdot}u = 0$. Now, by definition, $q{\cdot}v \neq 0$ for any proper left factor $v$ of $u$, so that $u \in S_q$. Now $0{\cdot}wu = 0$ and $wu \in [(P^*aP^*)^n]^*$ since $|p(u)|_{(0,a)} \equiv 0 \mod n$.

Conversely, assume that $w \in [(P^*aP^*)^n]^*S^{-1}$ and let $u \in S$ be a word such that $wu \in [(P^*aP^*)^n]^*$. Let $q = 0{\cdot}w$. Then $0{\cdot}wu = q{\cdot}u = 0$ and $q'.u = 0$ for every $q' \neq q$ since every word of $A^*$ induces a permutation on $Q$. It follows that $u \in S_q$. Now, by the definition of $S_q$,

$$|p(w)|_{(0,a)} = |p(wu)|_{(0,a)} \equiv 0 \mod n,$$

proving the claim. It follows that $h(L(\mathcal{A},(0,a),0,n)) \leq 1$.

For the general case we need a trick, which is just an extension of the following formula, in which $a$, $b$, and $c$ are distinct letters and $u$ is a word:

$$2|u|_a = |u|_{\{a,b\}} + |u|_{\{a,c\}} - |u|_{\{b,c\}}.$$

Thus, to compute $|u|_a$, it "suffices" to compute $|u|_{\{a,b\}}$, $|u|_{\{a,c\}}$ and $|u|_{\{b,c\}}$.

In our case, we want to find a similar formula to replace the computation of $|p(u)|_{(q_0,a)}$ by computations of numbers of the form $|p(u)|_{\{(q_1,a),(q_2,a),\ldots,(q_s,a)\}}$. For this purpose we introduce an abbreviation. If $S$ is a subset of $Q$, we put

$$|p(u)|_S = \sum_{q \in S} |p(u)|_{(q,a)}.$$

For every $j, j_1, \ldots, j_r \in \mathbb{Z}/p\mathbb{Z}$, we define the following subsets of $Q = (\mathbb{Z}/p\mathbb{Z})^r$:

$$H(j_1,\ldots,j_r;j) = \{(q_1,\ldots,q_r) \in Q \mid j_1q_1 + \ldots + j_rq_r = j\}$$

Then one can state

**Lemma 6.9** *For every word $u \in A^*$, the following equality holds:*

$$(p-1)p^{r-1}|p(u)|_{(q_0,a)} = (p-1)\sum_E |p(u)|_{H(1,j_2,\ldots,j_r;0)} - \sum_F |p(u)|_{H(0,j_2,\ldots,j_r;j)}$$

*where*

$$E = \{(j_2,\ldots,j_r) \in (\mathbb{Z}/p\mathbb{Z})^{r-1}\}$$

*and*

$$F = \{(j_2,\ldots,j_r,j) \in (\mathbb{Z}/p\mathbb{Z})^r \mid j \neq 0 \text{ and } (j_2,\ldots,j_r) \neq (0,\ldots,0)\}.$$

Before proving this lemma, let us write explicitly the formula when $p = 2$ and $r = 3$.

$$\begin{aligned}
4|p(u)|_{(0,0,0)} = &|p(u)|_{\{(0,0,0),(0,0,1),(0,1,0),(0,1,1)\}}\\
&+|p(u)|_{\{(0,0,0),(0,1,0),(1,0,1),(1,1,1)\}}\\
&+|p(u)|_{\{(0,0,0),(0,0,1),(1,1,0),(1,1,1)\}}\\
&+|p(u)|_{\{(0,0,0),(0,1,1),(1,0,1),(1,1,0)\}}\\
&-|p(u)|_{\{(0,0,1),(0,1,1),(1,0,1),(1,1,1)\}}\\
&-|p(u)|_{\{(0,1,0),(0,1,1),(1,1,0),(1,1,1)\}}\\
&-|p(u)|_{\{(0,0,1),(0,1,0),(1,0,1),(1,1,0)\}}
\end{aligned}$$

**Proof.** If we expand the right part of the formula by using the definition

$$|p(u)|_S = \sum_{q \in S} |p(u)|_{(q,a)},$$

we obtain a sum of the form $\sum_{q\in S} c_q |p(u)|_{(q,a)}$. The only thing to prove is that $c_{q_0} = (p-1)p^{r-1}$ and $c_q = 0$ if $q \neq q_0$. We first observe that

$$q_0 = (0, \ldots, 0) \in H(1, j_2, \ldots, j_r; 0) \text{ for } \textit{every } (j_2, \ldots, j_r) \in F$$

and that $q_0 \in H(0, j_2, \ldots, j_r; j)$ for *no* $(j, j_2, \ldots, j_r) \in F$. Therefore, $c_{q_0} = (p-1)\,\mathrm{Card}(E) = (p-1)p^{r-1}$.

Next assume that $q = (q_1, 0, \ldots, 0)$ for some $q_1 \neq 0$. Then

$$q \in H(1, j_2, \ldots, j_r; 0)$$

for *no* $(j_2, \ldots, j_r) \in E$ and $q \in H(0, j_2, \ldots, j_r; j)$ for *no* $(j, j_2, \ldots, j_r) \in F$. Thus $c_q = 0 - 0 = 0$ in this case.

Finally assume that $q = (q_1, q_2, \ldots, q_r)$ with $(q_2, \ldots, q_r) \neq (0, \ldots, 0)$. Then the equation in the unknown $j_2, \ldots, j_r$ defined by $q_1 + j_2q_2 + \ldots + j_rq_r = 0$ has exactly $p^{r-2}$ solutions in $E$, and the equation in the unknown $j, j_2, \ldots, j_r$ defined by $j_2q_2 + \ldots + j_rq_r = j$ has exactly $p^{r-2}(p-1)$ solutions in $F$. Thus $c_q = (p-1)p^{r-2} - (p-1)p^{r-2} = 0$ in this case also and this proves the lemma. $\square$

We return to the proof of Proposition 6.8. It follows from Lemma 6.9 and Proposition 6.2 that $L(\mathcal{A}, (q_0, a), k, n)$ is a boolean combination of languages of the form

$$\{u \in A^* \mid |p(u)|_{H(1,j_2,\ldots,j_r;0)} \equiv t \mod p^{r-1}n\}$$

or

$$\{u \in A^* \mid |p(u)|_{H(0,j_2,\ldots,j_r;j)} \equiv t \mod (p-1)p^{r-1}n\}$$

Thus the problem reduces to showing that

$$h(\{u \in A^* \mid |p(u)|_H \equiv k \mod n\}) \leq 1$$

where $H = H(j_1, j_2, \ldots, j_r; j)$ for some $(j_1, j_2, \ldots, j_r) \neq (0, \ldots, 0)$. Let $\gamma : Q \to \mathbb{Z}/p\mathbb{Z}$ be the function defined by

$$(q_1, q_2, \ldots, q_r)\gamma = j_1 q_1 + \ldots + j_r q_r$$

and let $\sim$ be the equivalence on $Q$ defined by $q \sim q'$ if and only if $q\gamma = q'\gamma$. Then $\sim$ is a congruence of the automaton $\mathcal{A}$. Indeed, if $a$ is a letter, $(q\cdot a)\gamma = (q + v_a)\gamma = q\gamma + v_a\gamma$. Therefore $q \sim q'$ implies $q\gamma + v_a\gamma = q'\gamma + v_a\gamma$, that is $q\cdot a \sim q'\cdot a$. One verifies immediately that the quotient automaton $\mathcal{A}/\!\sim$ is isomorphic to $\mathcal{A}' = (\mathbb{Z}/p\mathbb{Z}, A, ., 0)$ where, for every $a \in A$ and every $q \in \mathbb{Z}/p\mathbb{Z}$, $q\cdot a = q + v_a\gamma$. For instance, if $\mathcal{A}$ is the automaton represented in Figure 2, and if $H = \{(0,1), (1,0)\}$ is defined by $H = \{(q_1, q_2) \mid q_1 + q_2 = 1\}$, then $\mathcal{A}'$ will be represented as in Figure 3.

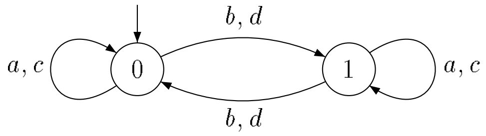



Figure 3:

Furthermore, $\gamma$ defines an automaton morphism from $\mathcal{A}$ onto $\mathcal{A}'$ that maps every element of $H$ onto $j$ (by definition of $H$). Therefore the following formula holds, where $p(u)$ (respectively $p'(u)$) denotes the path defined by $u$ in $\mathcal{A}$ (respectively in $\mathcal{A}'$):

$$|p(u)|_H = |p'(u)|_{(j,a)}$$

It follows that

$$\{u \in A^* \mid |p(u)|_H \equiv k \mod n\} = L(\mathcal{A}', (j, a), k, n).$$

But $\mathcal{A}'$ has $p$ states and hence $h(L(\mathcal{A}', (j, a), k, n)) \leq 1$ by the first part of the proof. $\square$

We now consider the case $p = 2$. Thus $\mathcal{A} = ((\mathbb{Z}/2\mathbb{Z})^r, A, q_0, \cdot)$ where $q_0 = (0, \ldots, 0)$ and, for every letter $a \in A$, there exists an $r$-tuple $v_a \in Q$ with $q\cdot a = q + v_a$ for all $q \in Q$. Then Proposition 6.8 can be slightly improved.

**Proposition 6.10** *Let $\mathcal{A} = ((\mathbb{Z}/2\mathbb{Z})^r, A, q_0, \cdot)$ be the automaton described above. Then for all the integers $k, n$ such that $0 \leq k < n$, for every letter $a \in A$, and for every state $q \in Q$, $h(L(\mathcal{A}, (q, a), k, n)) \leq 1$.*

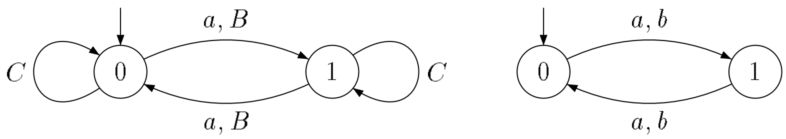
**Proof.** An argument similar to the end of the proof of Proposition 6.8 shows that one can assume $r = 1$, that is, $Q = \{0, 1\}$. Let $C$ be the set of all letters of $A$ inducing the identity on $Q$ and let $B = A \setminus (\{a\} \cup C)$. Let $\varphi : A^* \to \{a, b\}^*$ be the alphabetic morphism defined by

$$a\varphi = a, \quad b\varphi = \begin{cases} b & \text{if } b \in B \\ 1 & \text{if } b \in C. \end{cases}$$

Let $\mathcal{A}' = (\{0,1\}, \{a,b\}, \cdot, 0)$ be the automaton defined by the transitions

$$0\cdot a = 1, \quad 1\cdot a = 0, \quad 0\cdot b = 1 \text{ and } 1\cdot b = 0.$$

$\mathcal{A}$ and $\mathcal{A}'$ are represented in Figure 4.




$\mathcal{A}$ $\qquad\qquad$ $\mathcal{A}'$

Figure 4:

Let $p(u)$ (respectively $p'(u)$) be the path defined by a word $u$ in $\mathcal{A}$ (respectively $\mathcal{A}'$). Then, for every $u \in A^*$, $|p(u)|_{(0,a)} = |p'(u\varphi)|_{(0,a)}$, and hence,

$$L(\mathcal{A}, (0,a), k, n) = [L(\mathcal{A}', (0,a), k, n)]\varphi^{-1}.$$

Thus, by Corollary 4.7, it suffices to show that $h(L(\mathcal{A}', (0,a), k, n)) \leq 1$. Put $L = L(\mathcal{A}', (0,a), k, n)$ and $K = L \cap \{x \in \{a,b\}^* \mid 0\cdot x = 0\}$. We claim that

$$L = K\{1, b\}^{-1}.$$

Indeed, let $u \in L$ and let $q = 0\cdot u$. If $q = 0$, then $u \in K$. If $q = 1$, then $0\cdot ub = 1\cdot b = 0$ and $|p'(ub)|_{(0,a)} = |p'(u)|_{(0,a)}$. Thus $ub \in K$. Conversely, let $u$ be a word of $K\{1,b\}^{-1}$. Then $u \in K$ and hence $u \in L$ or $ub \in K$. Then $0\cdot ub = 0$ and $0\cdot u = 1$. It follows that $|p'(u)|_{(0,a)} = |p'(ub)|_{(0,a)}$ and $u \in L$, proving the claim.

Therefore, by Proposition 4.1, it suffices to show that $h(K) \leq 1$. Let $X$ be the prefix code $X = \{aa, ab, ba, bb\}$ and let $x \in X$. If $u \in X^*$, we denote by $|u|_x$ the number of occurrences of $x$ in the (unique) factorization of u as a product of words of $X$. Then

$$\begin{aligned} K &= \{u \in X^* \mid |u|_{ab} + |u|_{aa} \equiv k \mod n\} \\ &= \{u \in X^* \mid 2|u|_{ab} + 2|u|_{aa} \equiv 2k \mod 2n\}. \end{aligned}$$

Furthermore, for every $u \in X^*$, $|u|_{ab} + |u|_{ba} + 2|u|_{aa} = |u|_a$. It follows that

$$K = \{u \in X^* \mid |u|_a + |u|_{ab} - |u|_{ba} \equiv 2k \mod 2n\}.$$

Therefore, by Proposition 6.2, $K$ is a boolean combination of languages of the form

$$\{u \in X^* \mid |u|_a \equiv s \mod 2n\}$$

and of the form

$$\{u \in X^* \mid |u|_{ab} - |u|_{ba} \equiv s \mod 2n\}.$$

Languages of the first type are recognized by a commutative group and hence are of star height $\leq 1$ by Theorem 3.4. Let us now consider the languages of the second type. Let $\varphi : \{a,b\}^* \to \{a,b\}^*$ be the morphism defined by $a\varphi = ab$ and $b\varphi = ba$. Then $A^*\varphi = \{ab, ba\}^*$ is star-free (one can verify that its syntactic monoid is aperiodic and apply Theorem 3.1) and thus $\varphi$ is an injective star-free morphism by Proposition 4.4. Let

$$S = \{u \in \{a,b\}^* \mid |u|_a - |u|_b \equiv s \mod 2n\}.$$

Then $h(S) \leq 1$ by Theorem 3.4 and $h(S\varphi) \leq 1$ by Theorem 4.6. This concludes the proof, since $S\varphi = \{u \in X^* \mid |u|_{ab} - |u|_{ba} \equiv s \mod 2n\}$. □

# 7 Main results.

The results of Section 3 show that languages recognized by aperiodic monoids or by commutative groups are of star height $\leq 1$. The aim of this section is to prove similar results for some other varieties of monoids. Given a variety of monoids $\mathbf{V}$, we define the star height of $\mathbf{V}$ as the number

$$h(\mathbf{V}) = \max\{h(L) \mid L \text{ is recognized by a monoid of } \mathbf{V}\}$$

Thus $h(\mathbf{A}) = 0$ and $h(\mathbf{Gcom}) = 1$. Note that it is still an open problem to find a language (or a variety) of star height $> 1$! However one can prove a rather general result on $h(\mathbf{V})$. Recall that a monoid morphism $\varphi : M \to N$ is *aperiodic* if for every idempotent $e \in N$, $e\varphi^{-1}$ is an aperiodic subsemigroup of $M$. Given a variety of monoids $\mathbf{V}$, $\mathbf{A}^{-1}\mathbf{V}$ denotes the smallest variety containing all the monoids $M$ such that there exists an aperiodic morphism $\varphi : M \to N$ where $N \in \mathbf{V}$. Varieties of this form play an important role in semigroup theory (see [3, 8, 12] for more details). Then we have

**Proposition 7.1** *For every variety of monoids* $\mathbf{V}$, $h(\mathbf{V}) = h(\mathbf{A}^{-1}\mathbf{V}) = h(\mathbf{A} * \mathbf{V})$

**Proof.** Since $\mathbf{V} \subset \mathbf{A} * \mathbf{V} \subset \mathbf{A}^{-1}\mathbf{V}$, we have $h(\mathbf{V}) \leq h(\mathbf{A} * \mathbf{V})) \leq h(\mathbf{A}^{-1}\mathbf{V})$. Now by a theorem of [12], every language $L$ recognized by a monoid of $\mathbf{A}^{-1}\mathbf{V}$ is a boolean combination of languages of the form $L_0a_1L_1a_2\ldots a_kL_k$ where $k \geq 0$, $a_1, \ldots, a_k \in A$ and $L_0, \ldots, L_k$ are languages recognized by monoids of $\mathbf{V}$. Now since $h(L_0), \ldots, h(L_k) \leq h(\mathbf{V})$ by definition, $h(L) \leq h(\mathbf{V})$ and hence $h(\mathbf{A}^{-1}\mathbf{V}) \leq h(\mathbf{V})$ as required. □

Here is another general result which might be considered as a first step to compute $h(\mathbf{V})$.

**Proposition 7.2** *Let $\mathcal{F}$ be a class of finite monoids and let* $\mathbf{V}$ *be the variety of monoids generated by $\mathcal{F}$. If every language recognized by a monoid of $\mathcal{F}$ is of star-height $\leq n$, then* $h(\mathbf{V}) \leq n$.

**Proof.** Let $M \in \mathbf{V}$. Then $M$ divides a direct product $M_1 \times \ldots \times M_k$ of elements of $\mathcal{F}$. Let $L \subset A^*$ be a language recognized by $M$. Then since $M$ divides $M_1 \times \ldots \times M_k$, $L$ is also recognized by $M_1 \times \ldots \times M_k$. Therefore there exists a monoid morphism $\eta : A^* \to M_1 \times \ldots \times M_k$ and a subset $P$ of $M_1 \times \ldots \times M_k$ such that $L = P\eta^{-1}$. Since $L = \cup_{m \in P} m\eta^{-1}$, it suffices to prove that $h(m\eta^{-1}) \leq n$ for each $m \in P$. Denote by $\pi_i : M_1 \times \ldots \times M_k \to M_i$ the natural projection and set $m = (m_1, \ldots, m_k)$. Then $m\eta^{-1} = \cap_{1 \leq i \leq k} m_i \pi_i^{-1} \eta^{-1}$. But the language $L_i = m_i \pi_i^{-1} \eta^{-1}$ is recognized by $M_i$ so that $h(L_i) \leq n$ by assumption. Therefore $h(m\eta^{-1}) \leq n$ as required. □

Theorem 3.4 can be extended to the case of finite nilpotent groups of class 2.

**Theorem 7.3** *Every language recognized by a finite nilpotent group of class* 2 *is of star-height* $\leq 1$.

**Proof.** By Theorem 2.2, it suffices to show that the languages of the form $L(u, k, n)$, where $|u| \leq 2$ and $0 \leq k < n$, are of star height $\leq 1$. If $|u| \leq 1$, $L(u, k, n)$ is recognized by a commutative group (Theorem 2.1) and the result follows from Theorem 3.4. If $u = aa$ for some letter $a$, then $L(u, k, n)$ is also recognized by a commutative group. Indeed, since $\binom{v}{aa} = \binom{|v|_a}{2}$, we have $\binom{v}{aa} \equiv k \mod n$ if and only if there exists a positive integer $r$ such that $\binom{|v|_a}{2} \equiv k \mod n$ and $|v|_a \equiv r \mod 2n$. Therefore $L(aa, k, n)$ is a finite union of languages of the form $L(a, r, 2n)$.

Thus we may assume that $u = ab$ for some distinct letters $a$ and $b$. We define a deterministic complete automaton as follows. $\mathcal{A}_n = (\{0, 1, \ldots, n-1\}, A, \cdot, 0)$ where the transition function is given by

$$q \cdot a = q + 1 \mod n$$
$$q \cdot c = q \text{ if } c \neq a.$$

This automaton is represented in Figure 5.

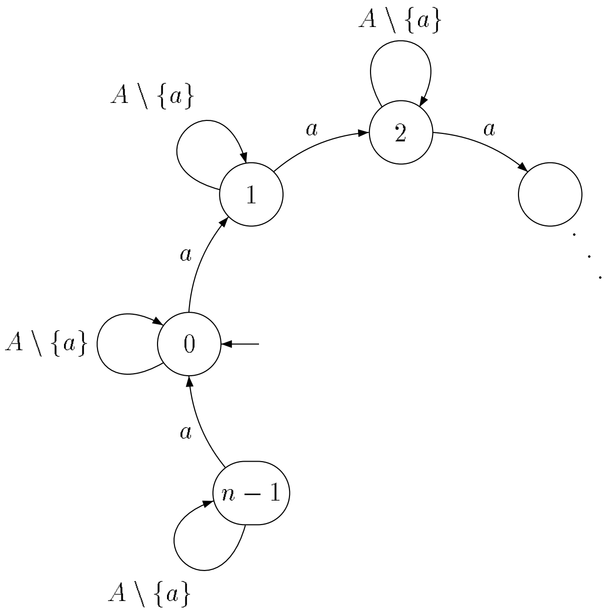



Figure 5:

We now observe that for every $v \in A^*$,

$$\binom{v}{ab} = \sum_{0 \leq i < n} i|p(v)|_{(i,b)} \mod k$$

where $p(v)$ denotes the path defined by $v$ in $\mathcal{A}_n$. Thus by Proposition 6.2, $L(ab, k, n)$ is a boolean combination of languages of the form $L(\mathcal{A}, (q, b), k_i, n)$, where $0 \leq k_i < n$. Now by Proposition 6.6, we have $h(L(\mathcal{A}, (q, b), k_i, n)) \leq 1$ and hence $L(ab, k, n) \leq 1$. $\square$

More generally, we have

**Theorem 7.4** *Let $a$ and $b$ be two letters of $A$. Then for every $i$, $j$, $k$, $n$ such that $0 \leq k < n$, $h(L(a^iba^j, k, n)) \leq 1$.*

**Proof.** Fix $i$, $j$ and $k$ and put $m = (\max(i, j))!$ and $N = mn$. Elementary arithmetic shows that if $s$, $t$ are positive integers such that $s \equiv t \mod N$, then

$$\binom{s}{i} \equiv \binom{t}{i} \mod n \quad \text{and} \quad \binom{s}{j} \equiv \binom{t}{j} \mod n.$$

Let $u \in A^*$. Then every factorization $u = xby$ defines exactly $\binom{|x|_a}{i}\binom{|y|_a}{j}$ occurrences of the subword $a^iba^j$ since this corresponds to the number of ways to take $i$ occurrences of $a$ in $x$ and $j$ occurrences of $a$ in $y$. Thus every $b$ that follows a number of $a$'s congruent to $s \mod N$ produces a number of subwords $a^i ba^j$ congruent to $\binom{s}{i}\binom{N+|u|_a-s}{j}$ modulo $n$.

Let $\mathcal{A}_N = (\{0, 1, \ldots, N-1\}, A, ., 0)$ be the automaton defined by

$$\begin{aligned} q{\cdot}a &= q+1 \mod N, \\ q{\cdot}c &= q \quad \text{for every letter } c \neq a. \end{aligned}$$

Thus, intuitively, $\mathcal{A}_N$ counts modulo $N$ the number of occurrences of $a$. Now we have

$$\binom{u}{a^i ba^j} \equiv \sum_{0 \le q < N} \binom{q}{i}\binom{N+|u|_a-q}{j} |p(u)|_{(q,b)} \mod n.$$

Thus $\binom{u}{a^i ba^j} \equiv k \mod n$ if and only if there exist an integer $s$ such that $0 \le s < N$ and integers $r_0, \ldots, r_{N-1} < N$ such that

(a) $s \equiv |u|_a \mod N$,

(b) $k \equiv \sum_{0 \le q < N} \binom{q}{i}\binom{N+s-q}{j} r_q \mod n$, and

(c) For $0 \le q < N$, $|p(u)|_{(q,b)} \equiv r_q \mod n$.

It follows by Proposition 6.2 that $L(a^i ba^j, k, n)$ is a boolean combination of languages of the form $L(a, s, N)$ and $L(\mathcal{A}_N, (q, a), r_q, n)$. But $h(L(a, s, N)) \le 1$ by Theorem 3.4 and $h(L(\mathcal{A}_N, (q, a), r_q, n)) \le 1$ by Proposition 6.7. Thus $h(L(a^i ba^j, k, n)) \le 1$. □

The next theorem is probably the most important result of this article. It shows in particular that the language $L(abc, 0, 2)$, which was considered as a possible candidate for star height 2 for the past ten years (Brzozowski, [1]), is in fact of star height one.

Recall that a number $n$ is square-free if it admits no square as a divisor.

**Theorem 7.5** *Let $a$, $b$ and $c$ be letters of $A$. If $n$ is a square-free number, then $h(L(abc, k, n)) \le 1$ for every $k$ such that $0 \le k < n$.*

**Proof.** If $a = c$, $b = c$ or $a = b$, the result follows from Theorem 7.4. Suppose now that $a$, $b$ and $c$ are three distinct letters and let $n = p_1 \cdots p_s$ be the decomposition of $n$ into prime numbers. Let, for $1 \le i \le s$, $k_i$ be a number such that $0 \le k_i < p_i$ and $k_i \equiv k \mod p_i$. Then by the Chinese Remainder Theorem, $x \equiv k \mod n$ if and only if $x \equiv k_i \mod p_i$ for $1 \le i \le s$, so that

$$L(abc, k, n) = \bigcap_{1 \le i \le s} L(abc, k_i, p_i).$$

Thus we may assume that $n = p$ is a prime number. The proof now mimics for the most part the proof of Theorem 7.4. Every factorization $u = xby$ defines $|x|_a|y|_c$ occurrences of the subword $abc$. Furthermore, we observe that $|y|_c =$

$|u|_c - |x|_c$. Let $\mathcal{A} = (Q, A, \cdot, q_0)$ be the automaton defined by $Q = (\mathbb{Z}/p\mathbb{Z})^2$, $q_0 = (0, 0)$ and

$$\begin{aligned}
&(q_1, q_2)\cdot a = (q_1 + 1, q_2),\\
&(q_1, q_2)\cdot c = (q_1, q_2 + 1),\\
&(q_1, q_2)\cdot d = (q_1, q_2) \text{ if } d \text{ is a letter different from } a \text{ and } c.
\end{aligned}$$

Thus, intuitively, $\mathcal{A}$ counts simultaneously modulo $p$ the number of occurrences of $a$'s and of $c$'s. Then we have

$$\binom{u}{abc} \equiv \sum_{q\in Q} q_1(|u|_c + p - q_2)|p(u)|_{(q,b)} \mod p.$$

Thus $\binom{u}{abc} \equiv k \mod p$ if and only if there exists an integer $s$ such that $0 \leq s < p$ and integers $r_q < n$ (for $q \in Q$) such that

(a) $s \equiv |u|_c \mod p$,

(b) $k \equiv \sum_{q\in Q} q_1(s + p - q_2)r_q \mod p$,

(c) For every $q \in Q$, $|p(u)|_{(q,b)} \equiv r_q \mod p$.

It follows that $L(abc, k, n)$ is a boolean combination of languages of the form $L(c, s, p)$ and $L(\mathcal{A}, (q, b), r_q, p)$. But $h(L, c, s, p) \leq 1$ by Theorem 3.4 and

$$h(L(\mathcal{A}, (q, b), r_q, p)) \leq 1$$

by Proposition 6.8. Thus $h(L(abc, k, n)) \leq 1$. □

We conclude this section with two results on the varieties of the form $\mathbf{V} * \mathbf{W}$.

**Theorem 7.6** *Every language recognized by a monoid of the variety* $\mathbf{Gcom} * (\mathbb{Z}/2\mathbb{Z})$ *is of star-height* $\leq 1$.

**Proof.** By definition, the variety $\mathbf{Gcom} * (\mathbb{Z}/2\mathbb{Z})$ is generated by the wreath products of the form $G \circ (\mathbb{Z}/2\mathbb{Z})^r$ where $G$ is a commutative group. Thus by Proposition 7.2, it suffices to show that every language recognized by such a wreath product is of star-height $\leq 1$.

Thus, let $\eta : A^* \to G \circ (\mathbb{Z}/2\mathbb{Z})^r$ be a morphism recognizing a language $L$. We denote by $\pi : G \circ (\mathbb{Z}/2\mathbb{Z})^r \to (\mathbb{Z}/2\mathbb{Z})^r$ the natural projection and we put $\varphi = \eta\pi$ and $B = (\mathbb{Z}/2\mathbb{Z})^r \times A$. By the wreath product principle, $L$ is a boolean combination of languages of the form $X \cap Y\sigma^{-1}$ where $X \subset A^*$ is recognized by $(\mathbb{Z}/2\mathbb{Z})^r$, $Y \subset B^*$ is recognized by $G$ and $\sigma : A^* \to B^*$ is the sequential function defined by

$$(a_1 \cdots a_r)\sigma = (1, a_1)(a_1\varphi, a_2) \cdots ((a_1 \cdots a_{r-1})\varphi, a_r).$$

Since $(\mathbb{Z}/2\mathbb{Z})^r$ is a commutative group, $h(X) \leq 1$ by Theorem 3.4 and it suffices to show that $h(Y\sigma^{-1}) \leq 1$. Since $Y$ is recognized by a commutative group, Theorem 2.1 shows that $Y$ is a boolean combination of languages of the form

$L(b,k,n)$ (where $b \in B$, and $0 \le k < n$). Since $\sigma^{-1}$ commutes with boolean operations, it is sufficient to prove that $h(L) \le 1$ where $L = \{u \in A^* \mid |u\sigma|_b \equiv k \mod n\}$. As we observed in Section 5, it is sufficient to show, for every $s$ such that $0 \le s < n$, that $h(L(\mathcal{A},(q,a),s,n) \le 1$ where $(q,a)$ is an arbitrary edge in the automaton $\mathcal{A}$ associated with $\sigma$. But $\mathcal{A} = ((\mathbb{Z}/2\mathbb{Z})^r, A, \cdot, 1)$ where the transition function is defined by $q\cdot a = q + (a\varphi)$. Thus $h(L) \le 1$ by Proposition 6.10. □

**Corollary 7.7** *Every language recognized by a group of order less that* 12 *is of star-height* $\le 1$.

**Proof.** Let $G$ be a finite group of order $n < 12$. If $n = p$ or $n = p^2$, where $p$ is prime, then $G$ is commutative, and if $n = p^3$, $G$ is nilpotent of class 2. Thus, if $n$ is different from 6 and 10, we may apply Theorem 3.4 or Theorem 7.3. If $n = 2m$, $G$ is either cyclic (and thus commutative) or equal to the dihedral group $D_m$. But $D_m$ can be decomposed as a semidirect product of the form $\mathbb{Z}/m\mathbb{Z} * \mathbb{Z}/2\mathbb{Z}$, and thus Theorem 7.6 can be applied. □

**Theorem 7.8** *Every language recognized by a monoid of the variety* $\mathbf{A} * \mathbf{Gcom} * \mathbf{A}$ *is of star-height* $\le 1$.

**Proof.** By Proposition 7.1, it suffices to show that every language recognized by a monoid of the variety $\mathbf{Gcom} * \mathbf{A}$ is of star height $\le 1$. By definition, $\mathbf{Gcom} * \mathbf{A}$ is generated by the wreath products of the form $G \circ M$ where $G$ is a commutative group and $M$ is an aperiodic monoid. Thus by Proposition 7.2, it suffices to show that every language recognized by such a wreath product $G \circ M$ is of star-height $\le 1$.

Thus, let $\eta : A^* \to G \circ M$ be a morphism recognizing a language $L$. We denote by $\pi : G \circ M \to M$ the natural projection and we put $\varphi = \eta\pi$ and $B = M \times A$. By the wreath product principle, $L$ is a boolean combination of languages of the form $X \cap Y\sigma^{-1}$ where $X \subset A^*$ is recognized by $M$, $Y \subset B^*$ is recognized by $G$ and $\sigma : A^* \to B^*$ is the sequential function defined by

$$(a_1 \cdots a_r)\sigma = (1, a_1)(a_1\varphi, a_2) \cdots ((a_1 \cdots a_{r-1})\varphi, a_r).$$

Since $M$ is aperiodic, $h(X) = 0$ by Theorem 3.1 and it suffices to show that $h(Y\sigma^{-1}) \le 1$. Since $Y$ is recognized by a commutative group, Theorem 2.1 shows that $Y$ is a boolean combination of languages of the form $L(b,k,n)$ (where $b \in B$, and $0 \le k < n$). Since $\sigma^{-1}$ commutes with boolean operations, it is sufficient to prove that $h(L) \le 1$ where $L = \{u \in A^* \mid |u\sigma|_b \equiv k \mod n\}$. As we observed in Section 5, it is sufficient to show, for every $s$ such that $0 \le s < n$, that $h(L(\mathcal{A},(q,a),s,n) \le 1$ where $(q,a)$ is an arbitrary edge in the automaton $\mathcal{A}$ associated with $s$. But $\mathcal{A} = (M, A, \cdot, 1)$ where the transition function is defined by $q\cdot a = q(a\varphi)$, so that the transition monoid of $\mathcal{A}$ is $M$, an *aperiodic* monoid. Therefore we can apply Proposition 6.7 to conclude the proof. □

# 8 Further results.

Unfortunately, it is even not known whether there exist languages of star-height greater than or equal to 2! A possible candidate is

$$L = (ab^*a \cup ba^*b(ab^*a)^*ba^*b)^*.$$

Notice that if $\mathcal{A}$ is the automaton represented on Figure 6,

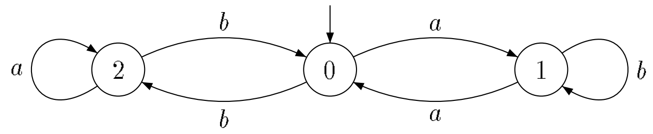



Figure 6:

then $L = \{u \in A^* \mid 0 \cdot u = 0 \text{ and } |p(u)|_{(0,b)} \equiv 0 \mod 2\}$.

More generally, good candidates can be found among the languages the syntactic monoid of which is a "sufficently complicated" finite group. The previous theorems suggest that languages of star-height $\leq n$ can be characterized through a property of their syntactic monoid. This hypothesis is explicitly stated in [5] but is very unlikely, unless every language is of star-height 0 or 1, according to the following result

**Theorem 8.1** *For every rational language $L$ of $A^*$, there exists a morphism $\varphi : A^* \to B^*$ and a rational language $K \subset B^*$ of restricted star-height $\leq 1$ such that $L = K\varphi^{-1}$.*

**Proof.** Let $\mathcal{A} = (Q, A, \cdot, 1, F)$ be the minimal automaton of $L$, and let $Q = \{1, 2, \ldots, n\}$. Let $B = A \cup \{c\}$ where $c$ is a letter not in $A$. Finally, let $\tau : \mathbb{N} \to \mathbb{N}$ be the function defined by

$$n\tau = 2^{n-1} - 1$$

(in fact the proof works with every function $\tau$ such that, for every $a$, $b$, $c$, $d$ in $\mathbb{N}$, $a\tau + b\tau = c\tau + d\tau$ implies $\{a, b\} = \{c, d\}$). Set

$$P = \{c^{i\tau} a c^{n\tau - (i \cdot a)\tau} \mid a \in A, i \in Q\},$$
$$S = \{c^{i\tau} \mid i \in F\} \quad \text{and} \quad R = P^*S.$$

Let $\varphi : A^* \to B^*$ be the morphism defined by $a\varphi = ac^{n\tau}$ for every $a \in A$. $R$ is, by construction, a language of restricted star-height 1 and furthermore $R\varphi^{-1} = L$. $\square$

In fact, Theorem 8.1 shows that languages of star-height $\leq n$ can even not be characterized through a property of their *pointed* syntactic monoid (if $\eta : A^* \to M$ is the syntactic morphism of $L$, the pair $(M, L\eta)$ is called the pointed monoid of $L$), unless every language is of star-height 0 or 1.

## Acknowledgements

A number of difficult computations on rational languages and semigroups were realized thanks to the remarkable C programs written by J.M. Champarnaud and G. Hansel [2]. The authors also thank the anonymous referees for several useful suggestions.

## References

[1] J. A. Brzozowski, Open problems about regular languages, Formal language theory, perspectives and open problems (R.V. Book editor), Academic Press, New York, 1980, 23–47.

[2] J.M. Champarnaud and G. Hansel, A computing package for automata and finite semigroups, *Journal of Symbolic Computation* **12**, (1991), 197–220.

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

[4] K. Hashiguchi, Representation theorems on regular languages, *J. Comput. System Sci.* **27**, (1983), 101–115.

[5] W.H. Henneman, Algebraic theory of automata, Ph. D. Dissertation, MIT (1971).

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

[7] R. McNaughton and S. Papert, *Counter-free Automata*, MIT-Press, Cambridge, Mass., 1971.

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

[9] J.E. Pin, H. Straubing and D. Thérien, New results on the generalized star-height problem, *STACS 89, Lecture Notes in Computer Science* **349**, (1989), 458–467.

[10] M.P. Schützenberger, On finite monoids having only trivial subgroups, *Information and Control*, **8**, (1965), 190–194.

[11] J. Stern, Complexity of some problems from the theory of automata, *Information and Computation*, **66**, (1985), 163-176.

[12] H. Straubing, Aperiodic homomorphisms and the concatenation product of recognizable sets, *J . Pure Appl. Algebra*, **15** (1979), 319-327.

[13] D. Thérien, Classification of regular languages by congruences, Ph. D. Thesis, Waterloo, 1980.

[14] D. Thérien, Subwords counting and nilpotent groups, in Combinatorics on Words, Progress and Perspectives, L. Cummings ed., Academic Press (1983), 293-306.

[15] W. Thomas, Remark on the Star-Height Problem, *Theoret. Comput. Sci.* **13**, 1981, 231-237.