Open problems on regular languages: an historical perspective

Laura Chaubard and Jean-Éric Pin[*]

May 14, 2006

Abstract

Operations on regular languages have been studied for fifty years, but several major problems remain wide open. This paper surveys the semigroup approach to these problems. We consider successively the star-height problem, the Straubing-Thérien's concatenation hierarchy and the shuffle operation. On the algebraic side, we present Eilenberg's variety theory and its successive improvements, including the recent notion of $\mathcal{C}$-variety.

Recall that a *language* is a subset of a finitely generated free monoid. The aim of this paper is to discuss various instances of the following general problem.

Problem. *Given a "basis" of languages, a set of operations and some rules to use them, describe the languages expressible from the basis by using the operations according to the rules.*

In practice, a basis of languages will consist of a set of very simple languages, such as the languages of the form $\{a\}$, where a is a letter of the alphabet. There are many possible choices for the operations, but we shall restrict ourselves to nine of them, that we now introduce.

1 Operations on languages

Let A be a finite alphabet and let A^* be the free monoid on A. Let us describe the operations we have in mind.

(1) *Boolean operations*, which comprise

  (a) finite *union* and finite *intersection* (these operations are also called the *positive Boolean operations*),

  (b) *complement* (denoted by $L \to L^c$).

(2) *Residual*: given a language L and a word u of A^*, $u^{-1}L = \{v \mid uv \in L\}$ and $Lu^{-1} = \{v \mid vu \in L\}$.

[*] LIAFA, Université Paris VII and CNRS, Case 7014, 2 Place Jussieu, 75251 Paris Cedex 05, France. Jean-Eric.Pin@liafa.jussieu.fr

(3) *Star*: L^* is the submonoid of A^* generated by L. Thus $L^* = \{u_1u_2 \cdots u_n \mid n \geqslant 0, u_1, \ldots, u_n \in L\}$.

(4) *Product*: the product of two languages L_1 and L_2 is the languages $L_1L_2 = \{u_1u_2 \mid u_1 \in L_1, u_2 \in L_2\}$.

(5) *Marked product*: given letters $a_1, \ldots, a_k$ of A and languages $L_0, L_1, \ldots, L_k$ of A^*, the marked product $L_0a_1L_1 \cdots a_kL_k$ is the language $\{u_0a_1u_1 \cdots a_ku_k \mid u_0 \in L_0, \ldots, u_k \in L_k\}$.

(6) *Shuffle product*. The shuffle of two words u and v of A^* is the set $u ш v$ of words of A^* of the form $u_1v_1 \cdots u_nv_n$, with $n \geqslant 0$, $u_1, \ldots, u_n, v_1, \ldots, v_n \in A^*$, $u_1 \cdots u_n = u$, $v_1 \cdots v_n = v$. For instance,

$$ab ш ba = \{abba, baab, abab, baba\}$$

The shuffle of two languages L_1 and L_2 of A^* is the set

$$L_1 ш L_2 = \bigcup_{u_1 \in L_1,\ u_2 \in L_2} u_1 ш u_2$$

(7) *Morphisms*. Let A and B be two alphabets, and let φ be a function from A into B^*. Then φ extends in a unique way into a morphism from A^* into B^*. If L is a language of A^*, $\varphi(L) = \{\varphi(u) \mid u \in L\}$ is a language of B^*.

(8) *Inverse morphisms*. If $\varphi\colon A^* \to B^*$ is a morphism and L is a language of B^*, then $\varphi^{-1}(L) = \{u \in A^* \mid \varphi(u) \in L\}$ is a language of A^*.

In our context, a *positive Boolean algebra* will be a class of languages closed under finite union and finite intersection. Since the empty language $\emptyset$ (resp. the full language A^*) can be considered as the union (resp. intersection) of an empty family of languages, they belong to all positive Boolean algebras. A Boolean algebra is a positive Boolean algebra closed under complement.

2 Rational and recognisable languages

Our first example is the class of *rational languages*. It is obtained by taking the languages $\{a\}$, for each letter a, as the basis and by allowing the use of only three operations, union, product and star, with no particular rules. If $A = \{a, b\}$, languages like A^*abaA^* or $(aba)^*ba \cup (bb(aa)^*ba)^*$ are rational.

Rational languages were characterised by Kleene in a seminal paper published in 1956 [13]. Kleene's theorem states that the rational languages are exactly the recognisable languages, that can be defined in (at least) three equivalent ways. We recall here two of these definitions, one relying on deterministic automata and one using finite monoids. A third possibility would be to make use of nondeterministic automata, but we shall not consider this approach in this paper.

A *finite automaton* is a quintuple $\mathcal{A} = (Q, A, E, q_0, F)$ where Q is a finite set (the set of *states*), A is an alphabet, E is a subset of $Q \times A \times Q$ (the set of *transitions*), q_0 is an element of Q (the *initial state*) and F is a subset of Q (the set of *final* states). Two transitions (p, a, q) and (p', a', q') are *consecutive* if $q = p'$. A *successful path* in $\mathcal{A}$ is a finite sequence of consecutive transitions starting in the initial state and ending in some final state

$$q_0 \xrightarrow{a_1} q_1 \xrightarrow{a_2} q_2 \ \cdots \ q_{n-1} \xrightarrow{a_n} q_n \in F$$

The word $a_1a_2 \cdots a_n$ is its *label.* The language recognised by $\mathcal{A}$ is the set of labels of all the successful paths in $\mathcal{A}$. A language is *recognisable* if it is recognised by some finite automaton.

A finite automaton is *deterministic* if for each state $p \in Q$ and each letter $a \in A$, there is at most one state q such that $(p, a, q) \in E$. This unique state q is denoted by $p \cdot a$. Thus each letter a induces a partial function $p \to p \cdot a$ from Q into itself. One can show that every recognisable language can be recognised by a deterministic automaton.

The definition involving monoids is more abstract. A monoid morphism $\varphi \colon A^* \to M$ *recognises* a language L of A^* if there is a subset P of M such that $L = \varphi^{-1}(P)$. By a slight abuse of language, we also say in this case that M recognises L.

It is not too difficult to show that the two definitions, by automata and by monoids, are equivalent. We can now reformulate Kleene's theorem as follows.

Theorem 2.1 (Kleene 1956) *For a language L, the following conditions are equivalent:*

(1) *L is rational,*

(2) *L is recognised by a finite monoid,*

(3) *L is recognised by a finite automaton.*

The term *regular* is also frequently usually used in the literature as an equivalent to *recognisable* or *rational.* It is important, however, to distinguish the latter two notions. First, both of them can be extended to arbitrary monoids, but they do not coincide in general. Secondly, depending on the problem, it might be more appropriate to take one definition or the other. Precise examples are given in the next paragraph.

A consequence of Kleene's theorem is that the class of recognisable languages is closed under the nine operations considered in Section 1. The importance of Kleene's theorem stems from the fact that some closure properties are transparent for rational expressions while others are much easier to prove using automata or monoids. For instance, it is straighforward to see that the class of rational languages is closed under union, (marked) product, star and morphisms. On the other hand, it is easy to see that recognisable languages are closed under complement, residuals, shuffle and inverse morphims. It is also possible, although slightly more difficult, to prove directly that recognisable languages are closed under (marked) product and star, but proving that rational languages are closed under complement without invoking Kleene's theorem is a real challenge. The skeptical reader may try to find a rational expression for the complement of the language $(((ab)^*aba)^*ba)^*$ to apprehend the difficulty of the problem.

The proof of Kleene's theorem is interesting for itself, since it provides an algorithm to convert a rational expression into a finite automaton and back. In the sequel, we shall meet several decidability problems of the form *decide whether a given regular language satisfies a certain property.* By Kleene's theorem, the solution of such a decision problem is independent of the representation chosen for the regular language, since descriptions by a rational expression, a finite automaton or a finite monoid can be translated one into another. However, the chosen representation has a strong influence on the complexity of the decision algorithms, a problem that we shall not address in this paper.

3 Star-height

In this section, we focus our attention on the star operation.

3.1 Star-free languages

The class of *star-free* languages is obtained by taking the languages $\{1\}$ and $\{a\}$, for each letter a, as the basic class $\mathcal{B}$ and by allowing Boolean operations and product. According to our general definition of a Boolean algebra, the languages $\emptyset$ and A^* are star-free. If $A = \{a, b\}$, the following languages are also star-free:

$$\begin{aligned} a^* &= (A^*bA^*)^c \\ (ab)^* &= (bA^* \cup A^*aaA^* \cup A^*bbA^* \cup A^*a)^c \end{aligned}$$

One can also show that the languages $\{ab, ba\}^*$ and $(a(ab)^*b)^*$ are star-free but that the languages $(aa)^*$ and $\{aba, b\}^*$ are not. Deciding whether a given rational language is star-free is a difficult problem, which was solved by Schützenberger in 1965.

Before stating this result, we need to introduce a few definitions. Let $\mathcal{A}$ be a finite deterministic automaton recognising a language L of A^*. A state q is called *accessible* if there exists a path from the initial state to q, and *coaccessible* if there is a path from q to some final state. By removing the states of $\mathcal{A}$ which are not simultaneously accessible and coaccessible, one obtains a *trim* automaton $\mathcal{B}$ that also recognises L. A further reduction consists in identifying two states p and q whenever, for every $u \in A^*$, $p \cdot u$ is final if and only if $q \cdot u$ is final. Performing this equivalence on the set of states of $\mathcal{B}$, one obtains a new automaton, called the *minimal automaton* of L, which also recognises L. For instance, the minimal automaton of the language $\{a, b\}^*aA^*b\{b, c\}^*$ is pictured in Figure 3.1.

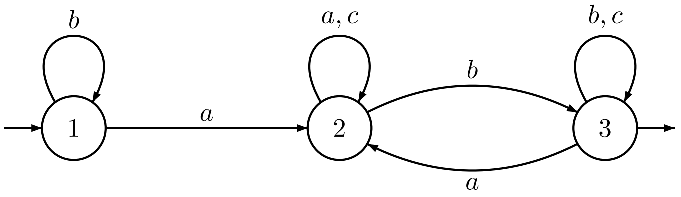


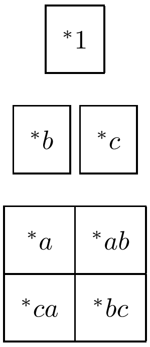
Figure 3.1: The minimal automaton of $\{a, b\}^*aA^*b\{b, c\}^*$.

The *syntactic monoid* of a language L can be defined in two equivalent ways. First, it is the transition monoid of the minimal automaton of L. Secondly, it is the quotient of A^* by the *syntactic congruence* of L, defined on A^* as follows: $u \sim_L v$ if and only if, for every $x, y \in A^*$,

$$xvy \in L \Leftrightarrow xuy \in L.$$

The syntactic monoid of the language $L = \{a, b\}^*aA^*b\{b, c\}^*$ and its $\mathcal{J}$-class structure are given below

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



Schützenberger [29] used the syntactic monoid to characterise the star-free languages. Recall that a finite monoid M is *aperiodic* if for each $x \in M$, there exists $n \geqslant 0$ such that $x^{n+1} = x^n$. Equivalently, a monoid is aperiodic if all the groups it contains are trivial, or if the Green's relation $\mathcal{H}$ is the equality.

Theorem 3.1 (Schützenberger 1965) *A language is star-free if and only if its syntactic monoid is finite and aperiodic.*

A consequence of Schützenberger's theorem is that one can effectively decide whether a given regular language is star-free. For instance, $\{a, b\}^*aA^*b\{b, c\}^*$ is star-free, since its syntactic monoid is aperiodic, but $(A^2)^*$ is not, since its syntactic monoid is the cyclic group of order 2.

3.2 The star-height problem

By Kleene's theorem, expressions built from letters by using Boolean operation, product and star represent regular languages. Such expressions are called *extended rational expressions*. The *star-height* of such an expression is the maximum number of nested stars occurring in the expression. For instance, the expression

$$(\{a, ba, abb\}^*bba \cap (aa\{a, ab\}^*))^c bbA^*$$

is of star-height one, while the expression

$$\Big(a(ba)^*abb\Big)^* bba \cap (aa\{a, ab\}^*))^c bbA^*$$

is of star-height two. The *star-height of a language* is the minimal star-height of an expression representing the language. In particular a language of star-height 0 is a star-free language.

We have seen that the language $(A^2)^*$ is not star-free. Since $(A^2)^*$ is an expression of star-height 1, this language has star-height exactly one. However, it is an open problem to know whether there are languages of star-height 2. We shall come back on this problem in Section 5.

4 Concatenation hierarchies

In this section, we introduce a hierarchy among star-free languages of A^*, known as the *Straubing-Thérien's hierarchy*, or *concatenation hierarchy*.[1] For historical reasons, this hierarchy is indexed by half-integers. The level 0 consists of the languages $\emptyset$ and A^*. The other levels are defined inductively as follows:

(1) the level $n + 1/2$ is the class of union of marked products of languages of level n;

(2) the level $n + 1$ is the class of Boolean combination of languages of marked products of level n.

We call the levels n (for some nonegative integer n) the *full levels* and the levels $n + 1/2$ the *half levels*.

It is not clear at first sight whether the Straubing-Thérien's hierarchy does not collapse, but this question was solved in 1978 by Brzozowski and Knast [6].

Theorem 4.1 (Brzozowski and Knast 1978) *The Straubing-Thérien's hierarchy is infinite.*

It is a major open problem on regular languages to know whether one can decide whether a given star-free language belongs to a given level.

Problem 2. *Given a half integer n and a star-free language L, decide whether L belongs to level n.*

One of the reasons why this problem is particularly appealing is its close connection with finite model theory, first explored by Büchi in the early sixties. Büchi's logic comprises a relation symbol $<$ and, for each letter $a \in A$ a predicate symbol $\mathbf{a}$. First order formulas are built in the usual way by using these symbols, the equality symbol, (first order) variables, Boolean connectives and quantifiers. Formal definitions can be found for instance in [32], but here we shall just present on an example how sentences are interpreted on finite words. The sentence

$$\varphi_1 = \exists x\ \exists y\ \big((x < y) \wedge (\mathbf{a}x) \wedge (\mathbf{b}y)\big),$$

can intuitively be interpreted on a word u by the English sentence "there exist two integers $x < y$ such that, in u, the letter in position x is an a and the letter in position y is a b". Therefore, the set of words satisfying φ_1 is $A^*aA^*bA^*$. McNaughton and Papert [15] showed that a language is first-order definable if and only if it is star-free. Thomas [32] (see also [16]) refined this result by showing that the concatenation hierarchy of star-free languages corresponds, level by level, to a hierarchy of first order formulas, the Σ_n-hierarchy. This hierarchy can be defined inductively as follows:

(1) Σ_0 consists of the quantifier-free formulas

(2) Σ_{n+1} consist of the formulas of the form $\exists x_1 \ldots \exists x_p \forall y_1 \ldots \forall y_q\ \varphi$, where $p, q \geqslant 0$ and φ is a Σ_n-formula.

(3) $\mathcal{B}\Sigma_n$ denotes the class of formulas that are Boolean combinations (that is, conjunctions of disjunctions) of Σ_n-formulas.

[1] A similar hierarchy, called the *dot-depth hierarchy* was previously introduced by Brzozowski, but the Straubing-Thérien's hierarchy is easier to define.

For instance, $\exists x_1 \exists x_2 \forall x_3 \forall x_4 \forall x_5 \exists x_6\, \varphi$, where φ is quantifier free, is in Σ_3. The next theorem summarizes the results of [15, 32, 16].

Theorem 4.2

(1) *A language is first-order definable if and only if it is star-free.*

(2) *A language is Σ_n-definable if and only if it is of level $n - 1/2$.*

(3) *A language is $\mathcal{B}\Sigma_n$-definable if and only if it is of level n.*

Thus deciding whether a language has level n is equivalent to a very natural problem in finite model theory. The first decidabilty result was obtained by I. Simon [30].

Theorem 4.3 (Simon 1972) *A language has level* 1 *if and only if its syntactic monoid is finite and $\mathcal{J}$-trivial.*

As in the case of star-free languages, the characterisation is given by a property of the syntactic monoid. This raises the question whether other families of regular languages can be described by an algebraic property of their syntactic monoid. The solution to this question was given by Eilenberg [10] in his variety theorem. We shall see in particular that the full levels of the concatenation hierarchy are varieties in Eilenberg's sense and thus can be described by some properties of their syntactic monoid. However, Eilenberg's theory does not apply to half levels, because they are not closed under complement. The solution, proposed in [20], consists of using the syntactic ordered monoid in place of the syntactic monoid. We briefly describe this extension before stating the variety theorem and its extended version.

First recall that an *ordered monoid* is a monoid equipped with an order $\leqslant$ compatible with the multiplication: $x \leqslant y$ implies $zx \leqslant zy$ and $xz \leqslant yz$.

We now give two equivalent definitions of the syntactic ordered monoid. We start with the algorithmic definition, which is probably easier to understand. Consider a minimal deterministic automaton $\mathcal{A} = (Q, A, \cdot, i, F)$. One defines a partial order $\leqslant$ on Q by $p \leqslant q$ if and only if, for each $u \in A^*$, $q \cdot u \in F \Rightarrow p \cdot u \in F$. For instance, for the automaton pictured in Figure 4.2, the partial order is $2 \leqslant 4$ and $1, 2, 3, 4 \leqslant 0$.

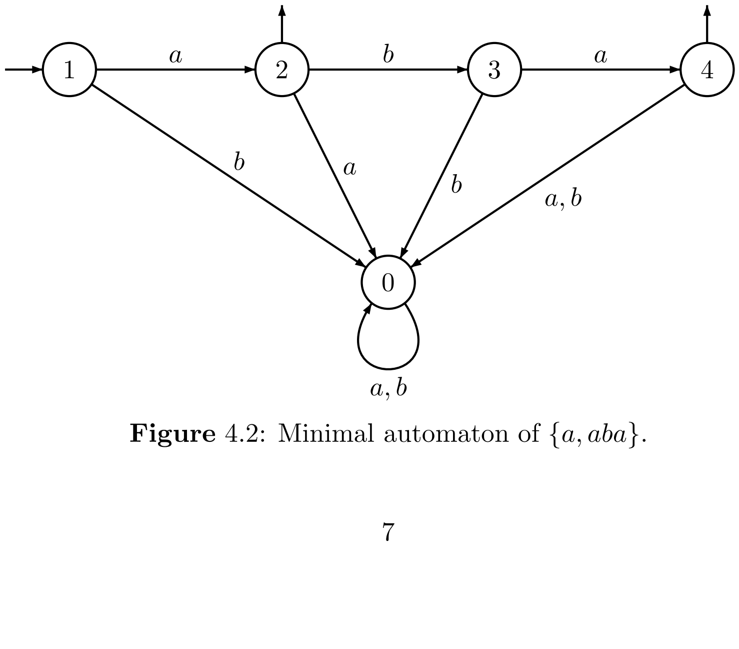


Figure 4.2: Minimal automaton of $\{a, aba\}$.

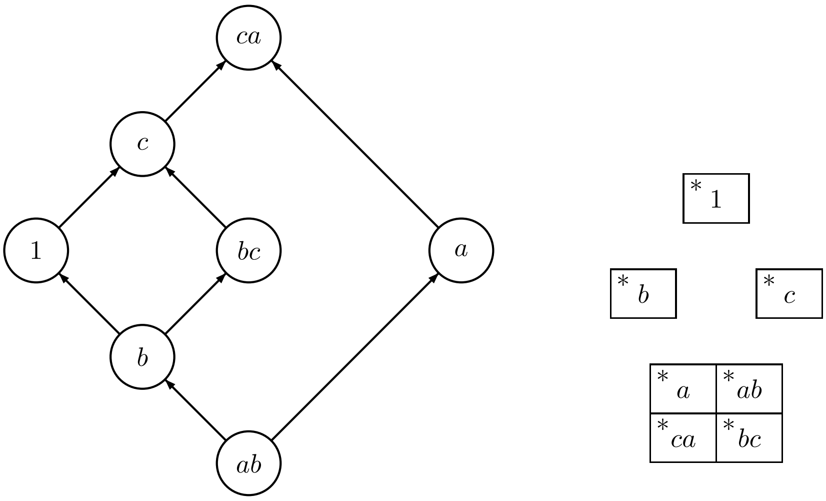
The *syntactic ordered monoid* of a language is the transition monoid of its minimal ordered automaton, ordered by $u \leqslant v$ if and only if for each $q \in Q$, $q \cdot u \leqslant q \cdot v$.

The second definition is more abstract. The *syntactic preorder* $\leqslant_L$ of a language L is defined as follows: $u \leqslant_L v$ iff, for every $x, y \in A^*$,

$$xvy \in L \Rightarrow xuy \in L$$

This preorder induces a partial order on the syntactic monoid of L, called the *syntactic order* of L. Thus the syntactic ordered monoid of L is equal to $(A^*/{\sim_L}, \leqslant_L/{\sim_L})$.

Example 4.1 The syntactic monoid and the syntactic order of the language $\{a, b\}^* a A^* b \{b, c\}^*$ are pictured below:



Thus ab is the smallest element in the syntactic order of L, and ca is the greatest.

A *variety of finite monoids* is a class of finite monoids closed under taking submonoids, quotients and finite products. Similarly, a *variety of finite ordered monoids* is a class of finite ordered monoids closed under taking ordered submonoids, quotients and finite products. Varieties of finite (ordered) *semigroups* are defined analogously. There is an abundant literature on varieties and we refer the reader to the books [1, 10, 19] for more details.

A convenient way to define varieties of finite monoids is to use identities. Let u, v be words of the free monoid A^*. A monoid M satisfies the *identity* $u = v$ if, for each morphism $\varphi\colon A^* \to M$, $\varphi(u) = \varphi(v)$. Similarly, an ordered monoid $(M, \leqslant)$ *satisfies the identity* $u \leqslant v$ if, for each morphism $\varphi : A^* \to M$, $\varphi(u) \leqslant \varphi(v)$. A variety of (ordered) monoids satisfies an identity if each of its monoids satisfies it.

The definition of an identity can be extended to *profinite identities*, which are formal equalities of the form $u = v$ (or $u \leqslant v$) where u and v are *profinite words*. We shall not attempt to define here profinite words nor profinite topology and the reader is referred to [2, 3, 33] for more details. We shall however define ω-terms, a special case of profinite words. An *ω-term* on an alphabet A is built from the letters of A using the usual concatenation product and the unary operator $x \to x^\omega$. Thus, if $A = \{a, b, c\}$, abc, a^ω and $((ab^\omega c)^\omega ab)^\omega$ are examples of ω-terms.

Let $\varphi : A^* \to M$ be a morphism from A^* into a finite monoid. The image $\varphi(t)$ of an ω-term t is defined recursively as follows. If t is a letter, then $\varphi(t)$ is already defined. If t and t' are ω-terms, then $\varphi(tt') = \varphi(t)\varphi(t')$. If $t = u^\omega$, then $\varphi(t)$ is the unique idempotent power of $\varphi(u)$.

Reiterman's theorem [28] ensures that a class of finite monoids is a variety if and only if it can be defined by a set of profinite identities. A similar result holds for varieties of ordered monoids. We refer to [3] for a detailed survey of this theory.

It is easy to prove directly that the class of finite (ordered) monoids (semigroups) satisfying a given set E of profinite identities is a variety of finite (ordered) monoids (semigroups), denoted by $[\![E]\!]$. Usually the context suffices to decide whether we are dealing with varieties of monoids or varieties of semigroups. For instance $[\![x^2 = x, xy = yx]\!]$ is the variety of finite idempotent and commutative monoids and $[\![x^\omega y x^\omega \leqslant x^\omega]\!]$ is the variety of all finite ordered semigroups S such that, for all $s \in S$ and $e \in E(S)$, $ese \leqslant e$.

A *positive variety of languages* is a class of recognisable languages $\mathcal{V}$ such that for any alphabets A and B,

(1) $\mathcal{V}(A^*)$ is a positive Boolean algebra,

(2) if $L \in \mathcal{V}(A^*)$ and $a \in A$ then $a^{-1}L, La^{-1} \in \mathcal{V}(A^*)$,

(3) if $\varphi\colon A^* \to B^*$ is a morphism, $L \in \mathcal{V}(B^*)$ implies $\varphi^{-1}(L) \in \mathcal{V}(A^*)$.

A *variety of languages* is a positive variety $\mathcal{V}$ such that, for each alphabet A, $\mathcal{V}(A^*)$ is closed under complement. We can now state the two variety theorems.

Theorem 4.4 (Eilenberg 1976) *Let* $\mathbf{V}$ *be a variety of finite monoids. For each alphabet A, let $\mathcal{V}(A^*)$ be the set of all languages of A^* whose syntactic monoid is in* $\mathbf{V}$. *Then $\mathcal{V}$ is a variety of languages. Further, the correspondence* $\mathbf{V} \to \mathcal{V}$ *is a bijection between varieties of finite monoids and varieties of languages.*

Theorem 4.5 (Pin 1995) *Let* $\mathbf{V}$ *be a variety of finite ordered monoids. For each alphabet A, let $\mathcal{V}(A^*)$ be the set of all languages of A^* whose syntactic ordered monoid is in* $\mathbf{V}$. *Then $\mathcal{V}$ is a positive variety of languages. Further, the correspondence* $\mathbf{V} \to \mathcal{V}$ *is a bijection between varieties of finite ordered monoids and positive varieties of languages.*

The next proposition shows that the variety approach is relevant for studying the concatenation hierarchy.

Proposition 4.6

(1) *The star-free languages form a variety of languages.*

(2) *Each full level of the concatenation hierarchy is a variety of languages.*

(3) *Each half level of the concatenation hierarchy is a positive variety of languages.*

We shall denote by $\mathbf{V}_n$ the variety of finite monoids corresponding to the languages of level n and by $\mathbf{V}_{n+1/2}$ the variety of ordered monoids corresponding to the languages of level $n + 1/2$.

Unfortunately, very few decidability results are known. It is obvious that a language has level 0 if and only if its syntactic monoid is trivial. The level 1/2 is also easy to study.

Theorem 4.7 (Pin-Weil 1995) *A language has level* $1/2$ *if and only if its ordered syntactic monoid* M *satisfies the identity* $x \leqslant 1$.

We already mentioned Simon's characterisation of languages of level 1. The decidability of level 3/2 was first proved by Arfi [4, 5] and the algebraic characterisation was found by Pin-Weil [24]. We need to introduce the Mal'cev product to state this result precisely.

Let $\mathbf{V}$ be variety of finite ordered semigroups and let M and N be two ordered monoids. A relational morphism $\tau : M \to N$ is a $\mathbf{V}$*-relational morphism* if, for every ordered subsemigroup T of N in $\mathbf{V}$, the ordered semigroup $\tau^{-1}(T)$ belongs to $\mathbf{V}$. Given a variety of finite monoids $\mathbf{W}$, the class of all ordered monoids M such that there exists a $\mathbf{V}$-relational morphism from M into an ordered monoid of $\mathbf{W}$ is a variety of ordered monoids, denoted by $\mathbf{V} \circledM \mathbf{W}$ and called the *Mal'cev product* of $\mathbf{V}$ and $\mathbf{W}$.

Theorem 4.8 (Pin-Weil 2001) *A language is of level* $3/2$ *if and only if its ordered syntactic monoid belongs to the Mal'cev product* $[\![x^\omega y x^\omega \leqslant x^\omega]\!] \circledM [\![x^2 = x, xy = yx]\!]$. *This condition is decidable.*

The decidability of level 2 is a major open problem in automata theory. An algebraic characterisation of $\mathbf{V}_2$ was given in [21], but it is not effective. Recall that a monoid M *divides* a monoid N if M is a quotient of a submonoid of N.

Theorem 4.9 (Pin-Straubing 1981) *A monoid belongs to* $\mathbf{V}_2$ *if and only if it divides a monoid of upper triangular Boolean matrices.*

Several partial results are known and a conjecture was proposed for the identities of $\mathbf{V}_2$, but its decidability is still open. See [27] for recent progress on this problem.

For the other levels, the decidability problem is also wide open. Pin and Weil [24, 26] established an algebraic connection between the varieties $\mathbf{V}_n$ and $\mathbf{V}_{n+1/2}$.

Theorem 4.10 (Pin-Weil 1995) *The variety* $\mathbf{V}_{n+1/2}$ *is equal to the Mal'cev product* $[\![x^\omega y x^\omega \leqslant x^\omega]\!] \circledM \mathbf{V}_n$.

Another result [25] describes, given the identities of a variety of finite monoids $\mathbf{V}$, a set of identities defining the variety $[\![x^\omega y x^\omega \leqslant x^\omega]\!] \circledM \mathbf{V}$.

Theorem 4.11 (Pin-Weil 1996) *The variety* $[\![x^\omega y x^\omega \leqslant x^\omega]\!] \circledM \mathbf{V}$ *is defined by the profinite identities* $u^\omega v u^\omega \leqslant u^\omega$, *where* u *and* v *are profinite words such that* $u = u^2$ *and* $u = v$ *are profinite identities of* $\mathbf{V}$.

These results illustrate the power of the algebraic approach, but do not suffice yet to show that if $\mathbf{V}_n$ is decidable, then $\mathbf{V}_{n+1/2}$ is decidable, except for $n = 1$.

5 Back to the star-height problem

Schützenberger's theorem gives a characterisation of the languages of star-height 0 and shows that they form a variety of languages. One may wonder whether this latter result also holds for the languages of star-height $\leqslant 1$. The answer to this question reduces to the existence of a language of star-height 2, as shown in [18]. Unfortunately, this problem is still open.

Theorem 5.1 (Pin 1978) *If the languages of star-height* $\leqslant 1$ *form a variety of languages, then there is no language of star-height* 2*.*

The closure properties of the languages of star-height $\leqslant n$ were analysed in [23]. Recall that a morphism between two free monoids is *length-preserving* if it maps each letter to a letter.

Theorem 5.2 (Pin, Straubing, Thérien 1989) *For each nonnegative integer n, the class of all languages of star-height $\leqslant n$ is closed under Boolean operations, residuals and inverse of length-preserving morphisms.*

Thus the languages of star-height $\leqslant n$ "almost" form a variety of languages. In fact, many other interesting classes of languages satisfy the two first conditions defining a variety of languages, but only a weak form of the third condition. Such examples include languages defined by fragments of first order logic or by temporal logic. Straubing [31] recently proposed a new extension of the notion of variety which covers these examples. A similar notion was introduced independently by Ésik and Ito [11].

Let $\mathcal{C}$ be a class of morphisms between free monoids, closed under composition and containing all length-preserving morphisms. Examples include the classes of all *length-preserving* morphisms, of all *length-multiplying* morphisms (morphisms such that, for some integer k, the image of any letter is a word of length k), all *non-erasing* morphisms (morphisms for which the image of each letter is a nonempty word), all *length-decreasing* morphisms (morphisms for which the image of each letter is either a letter or the empty word) and all morphisms.

A *positive $\mathcal{C}$-variety of languages* is a class $\mathcal{V}$ of recognisable languages satisfying the two first conditions defining a positive variety of languages and a third condition

(3′) if $\varphi\colon A^* \to B^*$ is a morphism in $\mathcal{C}$, $L \in \mathcal{V}(B^*)$ implies $\varphi^{-1}(L) \in \mathcal{V}(A^*)$.

A *$\mathcal{C}$-variety of languages* is a positive $\mathcal{C}$-variety of languages closed under complement. When $\mathcal{C}$ is the class of all (resp. length-preserving, length-multiplying, non-erasing, length-decreasing) morphisms, we use the term *all*-variety (resp. *lp*-variety, *lm*-variety, *ne*-variety, *de*-variety).

Theorem 5.2 gives an interesting example of *lp*-variety of languages.

Corollary 5.3 *For each $n \geqslant 0$, the languages of star-height $\leqslant n$ form an lp-variety of languages.*

The algebraic counterpart relies on a new syntactic invariant, the syntactic stamp. A *stamp* is a surjective morphism from A^* onto a finite monoid. The *syntactic stamp* of a regular language of A^* is the canonical morphism from A^* onto its syntactic monoid.

A stamp $\varphi : A^* \to M$ *$\mathcal{C}$-divides* a stamp $\psi : B^* \to N$ if there is a pair (f, η) (called a *$\mathcal{C}$-division*), where $f : A^* \to B^*$ is in $\mathcal{C}$, $\eta : N \to M$ is a partial surjective monoid morphism, and $\varphi = \eta \circ \psi \circ f$. If f is the identity on A^*, the pair (f, η) is simply called a division.

$$\begin{tikzcd}
A^* \arrow[rr, "f"] \arrow[d, "\varphi"'] & & B^* \arrow[d, "\psi"] \\
M & \mathrm{Im}(\psi \circ f) \subseteq \arrow[l, "\eta"'] & N
\end{tikzcd}$$

Figure 5.3: A division diagram.

The *product* of two stamps φ_1 and φ_2 is the stamp φ defined by $\varphi(a) = (\varphi_1(a), \varphi_2(a))$. A *$\mathcal{C}$-variety of stamps* is a class of stamps closed under $\mathcal{C}$-division and finite products.

Straubing's $\mathcal{C}$-variety theorem [31] can now be stated as follows.

Theorem 5.4 *Let $\mathbf{V}$ be a $\mathcal{C}$-variety of stamps. For each alphabet A, denote by $\mathcal{V}(A^*)$ the set of all languages of A^* whose syntactic stamp is in $\mathbf{V}$. Then $\mathcal{V}$ is a $\mathcal{C}$-variety of languages. Further, the correspondence $\mathbf{V} \to \mathcal{V}$ is a bijection between $\mathcal{C}$-varieties of stamps and $\mathcal{C}$-varieties of languages.*

The identity approach can be extended to $\mathcal{C}$-varieties of stamps as follows. Let u, v be two words of B^*. A stamp $\varphi : A^* \to M$ is said to *satisfy the $\mathcal{C}$-identity $u = v$* if, for every $\mathcal{C}$-morphism $f\colon B^* \to A^*$, $\varphi \circ f(u) = \varphi \circ f(v)$. If M is ordered, we say that φ satisfies the $\mathcal{C}$-identity $u \leqslant v$ if, for every $\mathcal{C}$-morphism $f\colon B^* \to A^*$, $\varphi \circ f(u) \leqslant \varphi \circ f(v)$. By extension, we say that a language satisfies an identity if its syntactic stamp satisfies this identity.

Example 5.1 Let $\varphi : A^* \to M$ be a stamp. Consider the identity

$$xyx = x \tag{1}$$

If $\mathcal{C}$ is the class of all morphisms, φ satisfies (1) if and only if, for all $x, y \in A^*$, $\varphi(xyx) = \varphi(x)$. Now, if $\mathcal{C}$ is the class of length-preserving morphisms, φ satisfies (1) if and only if, for all $x, y \in A$, $\varphi(xyx) = \varphi(x)$. If $\mathcal{C}$ is the class of length-multiplying morphisms, φ satisfies (1) if and only if, for each $k \geqslant 0$ and for all $x, y \in A^k$, $\varphi(xyx) = \varphi(x)$.

The definition of identities can be extended to profinite identities to obtain a generalisation of Reiterman's theorem to $\mathcal{C}$-varieties [14, 22].

It follows from the previous results that the star-height problem amounts to showing that the *lp*-varieties of stamps corresponding to the languages of star-height $\leqslant n$ are decidable. But even if these varieties of stamps cannot be characterized precisely, one can still hope to find some identity satisfied by all languages of star-height $\leqslant 1$. It would then suffice to find a regular language not satisfying this identity to have an example of a language of star-height > 1.

For recent developments about $\mathcal{C}$-varieties, we refer the reader to the papers [8, 9, 22].

6 Shuffle product

Introducing the shuffle product into the picture leads to several interesting questions. First, what are the varieties of languages closed under shuffle? The commutative varieties of languages closed under shuffle were characterised by Perrot [17]: they correspond to the varieties of commutative monoids whose groups belong to a given variety of commutative groups. The variety of all rational languages is also closed under shuffle. Are there other examples? Esik and Simon [12] answered this question negatively. Let us say that a variety of languages is *proper* if it is not equal to the variety of all rational languages.

Theorem 6.1 (Esik-Simon 1998) *The variety of commutative languages is the largest proper variety of languages closed under shuffle.*

Is there a similar result for positive varieties of languages? That is, is there a largest proper positive variety of languages closed under shuffle? The answer was given in [7].

Theorem 6.2 (Cano Gómez, Pin 2004) *There is a largest positive variety not containing* $(ab)^*$. *It is also the largest proper positive variety closed under length preserving morphisms and the largest proper positive variety closed under shuffle.*

A characterisation of the corresponding variety of ordered monoids $\mathbf{W}$ was given in the same paper.

Theorem 6.3 (Cano Gómez, Pin 2004) *An ordered monoid belongs to* $\mathbf{W}$ *if and only if, for every pair* (a, b) *of mutually inverse elements, and for every element* z *of the minimal ideal of the submonoid generated by* a *and* b, $(abzab)^\omega \leqslant ab$. *In particular* $\mathbf{W}$ *is decidable.*

It would be interesting to know whether a similar result holds for *lp*-varieties of languages: is there a largest proper *lp*-variety of languages closed under shuffle? Is there a largest proper positive *lp*-variety of languages closed under shuffle?

7 Conclusion

The successive improvements over Eilenberg's variety theory have considerably enlarged the scope of the algebraic approach to the study of regular languages. It has been applied succesfully to a large range of applications, including logic and finite model theory, circuit complexity, abstract complexity, communication complexity, infinite words and other structures. However, several exciting problems remain unsolved and we would like to encourage the semigroup community to work on these questions.

References

[1] J. Almeida, *Finite semigroups and universal algebra. Series in Algebra*, vol. 3, World Scientific, Singapore, 1994.

[2] J. Almeida, Profinite semigroups and applications, in *Structural theory of automata, semigroups, and universal algebra, Proceedings of the NATO Advanced Study Institute, Montreal, Quebec, Canada, July 7-18, 2003.*, V. B. e. a. Kudryavtsev (ed.), pp. 1–45, *NATO Science Series II: Mathematics, Physics and Chemistry 207*, Kluwer Academic Publishers, 2005.

[3] J. Almeida and P. Weil, Relatively free profinite monoids: an introduction and examples, in *NATO Advanced Study Institute Semigroups, Formal Languages and Groups*, J. Fountain (ed.), vol. 466, pp. 73–117, Kluwer Academic Publishers, 1995.

[4] M. Arfi, Polynomial operations on rational languages, in *STACS 87 (Passau, 1987)*, pp. 198–206, *Lecture Notes in Comput. Sci.* vol. 247, Springer, Berlin, 1987.

[5] M. Arfi, Opérations polynomiales et hiérarchies de concaténation, *Theoret. Comput. Sci.* **91**,1 (1991), 71–84.

[6] J. A. Brzozowski and R. Knast, The dot-depth hierarchy of star-free languages is infinite, *J. Comput. System Sci.* **16**,1 (1978), 37–55.

[7] A. Cano Gómez and J.-É. Pin, Shuffle on positive varieties of languages, *Theoret. Comput. Sci.* **312**,2-3 (2004), 433–461.

[8] L. Chaubard, Actions and wreath products of $\mathcal{C}$-varieties, in *LATIN'06 (Valdivia, 2006)*, Berlin, 2006, pp. 274–285, *Lecture Notes in Comput. Sci.* vol. 3887, Springer.

[9] L. Chaubard, J.-É. Pin and H. Straubing, Actions, Wreath Products of $\mathcal{C}$-varieties and Concatenation Product, *Theoret. Comput. Sci.*, 2006. to appear.

[10] S. Eilenberg, *Automata, languages, and machines. Vol. B*, Academic Press [Harcourt Brace Jovanovich Publishers], New York, 1976. With two chapters ("Depth decomposition theorem" and "Complexity of semigroups and morphisms") by Bret Tilson, Pure and Applied Mathematics, Vol. 59.

[11] Z. Ésik and M. Ito, Temporal Logic with Cyclic Counting and the Degree of Aperiodicity of Finite Automata, *Acta Cybernetica* **16** (2003), 1–28.

[12] Z. Ésik and I. Simon, Modeling Literal Morphisms by Shuffle, *Semigroup Forum* **56** (1998), 225–227.

[13] S. C. Kleene, Representation of Events in nerve nets and finite automata, in *Automata Studies*, C. Shannon and J. McCarthy (ed.), Princeton, New Jersey, 1956, pp. 3–42, Princeton University Press.

[14] M. Kunc, Equational description of pseudovarieties of homomorphisms, *Theoretical Informatics and Applications* **37** (2003), 243–254.

[15] R. McNaughton and S. Papert, *Counter-free automata*, The M.I.T. Press, Cambridge, Mass.-London, 1971. With an appendix by William Henneman, M.I.T. Research Monograph, No. 65.

[16] D. Perrin and J.-E. Pin, First-order logic and star-free sets, *J. Comput. System Sci.* **32**,3 (1986), 393–406.

[17] J.-F. Perrot, Variétés de langages et operations, *Theoret. Comput. Sci.* **7** (1978), 197–210.

[18] J.-É. Pin, Sur le monoïde de L^* lorsque L est un langage fini, *Theoret. Comput. Sci.* **7** (1978), 211–215.

[19] J.-É. Pin, *Varieties of formal languages*, North Oxford, London and Plenum, New-York, 1986. (Traduction de Variétés de langages formels).

[20] J.-É. Pin, A variety theorem without complementation, *Russian Mathematics (Izvestija vuzov.Matematika)* **39** (1995), 80–90.

[21] J.-E. Pin and H. Straubing, Monoids of upper triangular matrices, in *Semigroups (Szeged, 1981)*, pp. 259–272, *Colloq. Math. Soc. János Bolyai* vol. 39, North-Holland, Amsterdam, 1985.

[22] J.-E. Pin and H. Straubing, Some results on $\mathcal{C}$-varieties, *Theoret. Informatics Appl.* **39** (2005), 239–262.

[23] J.-É. Pin, H. Straubing and D. Thérien, Some results on the generalized star-height problem, *Information and Computation* **101** (1992), 219–250.

[24] J.-É. Pin and P. Weil, Polynomial closure and unambiguous product, in *22th ICALP*, Berlin, 1995, pp. 348–359, *Lect. Notes Comp. Sci.* n° 944, Springer.

[25] J.-É. Pin and P. Weil, Profinite semigroups, Mal'cev products and identities, *J. of Algebra* **182** (1996), 604–626.

[26] J.-É. Pin and P. Weil, Polynomial closure and unambiguous product, *Theory Comput. Systems* **30** (1997), 1–39.

[27] J.-É. Pin and P. Weil, A conjecture on the concatenation product, *ITA* **35** (2001), 597–618.

[28] J. Reiterman, The Birkhoff theorem for finite algebras, *Algebra Universalis* **14**,1 (1982), 1–10.

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

[30] I. Simon, Piecewise testable events, in *Proc. 2nd GI Conf.*, H. Brackage (ed.), pp. 214–222, *Lecture Notes in Comp. Sci.* vol. 33, Springer Verlag, Berlin, Heidelberg, New York, 1975.

[31] H. Straubing, On logical descriptions of regular languages, in *LATIN 2002*, Berlin, 2002, pp. 528–538, *Lect. Notes Comp. Sci.* n° 2286, Springer.

[32] W. Thomas, Classifying regular events in symbolic logic, *J. Comput. System Sci.* **25**,3 (1982), 360–376.

[33] P. Weil, Profinite methods in semigroup theory, *Int. J. Alg. Comput.* **12** (2002), 137–178.