Equational descriptions of languages

Jean-Eric Pin

Equational descriptions of languages*

Jean-Éric Pin[1]

May 28, 2012

Abstract

This paper is a survey on the equational descriptions of languages. The first part is devoted to Birkhoff's and Reiterman's theorems on equational descriptions of varieties. Eilenberg's variety theorem and its successive generalizations form the second part. The more recent results on equational descriptions of lattices of languages are presented in the third part of the paper.

Equations have been used for a long time in mathematics to provide a concise description of various mathematical objects. This article roughly follows a historical approach to present such equational descriptions for formal languages, ranging over a period of 45 years: from Schützenberger's characterization of star-free languages [36] to the following recent result of [18]: *Every lattice of languages admits an equational description.*

This evolution was made possible by a gradual abstraction of the notion of equation. The story really starts in 1935 with Birkhoff's theorem on equational classes [6]. It holds for any kind of universal algebra, but I will present it only for monoids.

1 Varieties and identities

A *Birkhoff variety of monoids* is a class of (possibly infinite) monoids which is closed under taking submonoids, homomorphic images (also called quotients) and arbitrary direct products.

Birkhoff's theorem states that a class of monoids is a Birkhoff variety if and only if it can be defined by a set of identities. Formally, an *identity* is an equality between two words of the free monoid Σ^* on a countable alphabet Σ. A monoid M *satisfies the identity* $u = v$ if and only if $\varphi(u) = \varphi(v)$ for every morphism $\varphi : \Sigma^* \to M$. For instance, the identity $xy = yx$ defines the variety of commutative monoids and $x = x^2$ defines the variety of idempotent monoids.

It is easy to see that if E is a set of identities, the class of monoids satisfying all the identities of E is a variety, denoted $[\![E]\!]$, and called the *Birkhoff variety defined by* E. The difficult part of Birkhoff's theorem is to prove that the converse also holds: for each Birkhoff variety $\mathbf{V}$, there exists a set of identities E such that $\mathbf{V} = [\![E]\!]$. Note that the set E might be infinite. For instance, if

[1]LIAFA, CNRS and Univ. Paris-Diderot, Case 7014, 75205 Paris Cedex 13, France.

*The author acknowledges support from the project ANR 2010 BLAN 0202 02 FREC.

$M = \{1, a, b, ab, ba, 0\}$ with $aba = a$, $bab = b$ and $aa = bb = 0$, then the Birkhoff variety generated by M is not finitely based: no finite set of identities suffices to define it [21].

An important step forward was Reiterman's customization of Birkhoff's theorem for varieties of finite monoids.

A *variety of finite monoids* is a class of finite monoids which is closed under taking submonoids, quotients and finite direct products. For instance, the class of finite aperiodic[1] monoids forms a variety. Similarly, the class of $\mathcal{J}$-trivial[2] monoids forms a variety. The class of all finite groups also forms a variety of finite monoids.

It is a natural question to ask whether varieties of finite monoids can also be defined by identities. The problem was solved by several authors but the most satisfactory answer is due to Reiterman [34] (see also [3]). A little bit of topology is required to state this theorem.

A finite monoid M *separates* two words u and v of A^* if there is a monoid morphism $\varphi : A^* \to M$ such that $\varphi(u) \neq \varphi(v)$. We set

$$r(u, v) = \min\{\mathrm{Card}(M) \mid \ M \text{ is a finite monoid that separates } u \text{ and } v \ \}$$

and $d(u, v) = 2^{-r(u,v)}$, with the usual conventions $\min \emptyset = +\infty$ and $2^{-\infty} = 0$. Then d is a *metric* on A^* and the completion of A^* for this metric is denoted by $\widehat{A^*}$. The (concatenation) product on A^* can be extended by continuity to $\widehat{A^*}$, making $\widehat{A^*}$ a compact topological monoid, called the *free profinite monoid*. Its elements are called *profinite words*. It is not so easy to give examples of profinite words which are not words, but here is one. In a compact monoid, the smallest closed subsemigroup containing a given element x has a unique idempotent, denoted x^ω. This is true in particular in a finite monoid and in the free profinite monoid. Thus if x is a (profinite) word, so is x^ω. In fact, one can show that x^ω is the limit of the converging sequence $x^{n!}$. More details can be found in [1, 27].

Every finite monoid M can be considered a discrete metric space for the metric d, defined by $d(x, y) = 0$ if $x = y$, and $d(x, y) = 1$ otherwise. Now, every morphism φ from A^* into a finite monoid is uniformly continuous and therefore can be extended (in a unique way) into a uniformly continuous morphism $\widehat{\varphi}$ from $\widehat{A^*}$ to M. It follows for instance that if x is a profinite word, then $\widehat{\varphi}(x^\omega) = \widehat{\varphi}(x)^\omega$.

Let u and v be two profinite words of $\widehat{A^*}$. A monoid M *satisfies the profinite identity* $u = v$ if, for each monoid morphism $\varphi : A^* \to M$, one has $\widehat{\varphi}(u) = \widehat{\varphi}(v)$.

Reiterman's theorem states that a class of finite monoids is a variety if and only if it can be defined by a set of profinite identities. Thus the core of Reiterman's theorem is a topological extension of Birkhoff's theorem. For instance, the variety of aperiodic monoids is defined by the identity $x^{\omega+1} = x^\omega$. The variety of $\mathcal{J}$-trivial monoids is defined by the identities $x^{\omega+1} = x^\omega$ and $(xy)^\omega = (yx)^\omega$. The variety of finite groups is defined by the identity $x^\omega = 1$.

[1] An element x of a monoid is *aperiodic* if there exists an integer $n > 0$ such that $x^{n+1} = x^n$. A finite monoid is *aperiodic* if all its elements are aperiodic.

[2] In a monoid, the $\leqslant_{\mathcal{J}}$ preorder is defined by $u \leqslant v$ if $u = xvy$ for some $x, y \in M$. A monoid is $\mathcal{J}$-trivial when this preorder is an order.

2 Eilenberg's variety theorem

Let us now come back to languages. A regular language is said to *satisfy a profinite identity* if its syntactic monoid satisfies this identity. The first example of profinite identities defining a class of regular languages is a consequence of Schützenberger's theorem on star-free languages [36].

Theorem 2.1 *Let L be a regular language. The following conditions are equivalent:*

(1) *L is star-free,*
(2) *the syntactic monoid of L is aperiodic,*
(3) *L satisfies the identity $x^{\omega+1} = x^\omega$.*

Another famous example is Simon's theorem on piecewise testable languages. Let us call *simple* a language of the form

$$A^* a_1 A^* a_2 A^* \cdots a_k A^*$$

where $k \geqslant 0$ and $a_1, a_2, \ldots, a_k \in A$. A language is *piecewise testable* if it is a Boolean combination of simple languages. Simon proved the following result [37].

Theorem 2.2 *Let L be a regular language. The following conditions are equivalent:*

(1) *L is piecewise testable,*
(2) *the syntactic monoid of L is $\mathcal{J}$-trivial,*
(3) *L satisfies the identities $x^{\omega+1} = x^\omega$ and $(xy)^\omega = (yx)^\omega$.*

These theorems and similar results led Eilenberg to his celebrated variety theorem. Let me first recall some standard definitions.

A *lattice of languages* is a set $\mathcal{L}$ of languages of A^* containing $\emptyset$ and A^* and closed under finite unions and finite intersections. It is *closed under quotients*[3] if, for each $L \in \mathcal{L}$ and $u \in A^*$, the languages $u^{-1}L$ and Lu^{-1} are also in $\mathcal{L}$. A lattice is a *Boolean algebra* if it is closed under complement.

A *class of languages* $\mathcal{C}$ associates with each finite alphabet A a set $\mathcal{C}(A^*)$ of regular languages of A^*. It is *closed under inverses of morphisms* if, for each morphism $\varphi\colon A^* \to B^*$, the condition $L \in \mathcal{C}(B^*)$ implies $\varphi^{-1}(L) \in \mathcal{C}(A^*)$.

A *variety of languages* is a class of languages $\mathcal{V}$ such that:

(1) for each alphabet A, $\mathcal{V}(A^*)$ is a Boolean algebra of languages closed under quotients,
(2) $\mathcal{V}$ is closed under inverses of morphisms.

Eilenberg's variety theorem [13] gives a bijective correspondence between varieties of finite monoids and varieties of languages.

Theorem 2.3 *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 monoids and varieties of languages.*

[3] Recall that $u^{-1}L = \{x \in A^* \mid ux \in L\}$ and $Lu^{-1} = \{x \in A^* \mid xu \in L\}$.

The conjunction of Eilenberg's and Reiterman's theorems gives the following result.

Corollary 2.4 *Any variety of languages can be defined by a set of profinite identities.*

For instance, the variety of star-free languages is defined by the profinite identity $x^{\omega+1} = x^\omega$. The variety of piecewise testable languages is defined by the two identities $x^{\omega+1} = x^\omega$ and $(xy)^\omega = (yx)^\omega$.

The Eilenberg-Reiterman theorem is also useful to analyse the expressive power of various fragments of first order logic interpreted on finite words. One associates to each nonempty word $u = a_0a_1 \dots a_{|u|-1}$ over the alphabet A a relational structure

$$\mathfrak{M}_u = \{(0, 1, \dots, |u| - 1), <, (\mathbf{a})_{a\in A}\}$$

where $<$ is the usual order on the domain and $\mathbf{a}$ is a predicate giving the positions i such that $a_i = a$. For instance, if $u = abbaaba$, then $\mathbf{a} = \{0, 3, 4, 6\}$ and $\mathbf{b} = \{1, 2, 5\}$. Given a formula φ, the language defined by φ is $L(\varphi) = \{u \in A^+ \mid \mathfrak{M}_u \text{ satisfies } \varphi\}$. Since languages may contain the empty word, we make the convention that a language L of A^* is defined by φ if $L(\varphi) = L \cap A^+$.

McNaughton and Papert [20] showed that a language is first-order definable (in the signature $\{<, (\mathbf{a})_{a\in A}\}$) if and only if it is star-free. Thomas [40] (see also [22]) refined this result by showing that the concatenation hierarchy of star-free languages is, level by level, in correspondence with the Σ_n-hierarchy of first order formulas. In particular, he showed that the Boolean closure of Σ_1-formulas, denoted by $\mathcal{B}\mathbf{\Sigma}_1[<]$, captures exactly the piecewise testable languages. These results provide an equational description of $\mathbf{FO}[<]$ and $\mathcal{B}\mathbf{\Sigma}_1[<]$ which can be formulated as follows:

Theorem 2.5 *The following formulas hold:* $\mathbf{FO}[<] = [\![x^{\omega+1} = x^\omega]\!]$ *and* $\mathcal{B}\mathbf{\Sigma}_1[<] = [\![x^{\omega+1} = x^\omega, (xy)^\omega = (yx)^\omega]\!]$.

3 Bringing order...

Originally motivated by an attempt to give a purely algebraic characterization of the regular open sets of the pro-group topology [24, 26], the author proposed an extension to Eilenberg's variety theorem [25] that covers classes of languages that are not necessarily closed under complement. The idea is to use ordered monoids instead of monoids. Let us briefly recall the main definitions.

An *ordered monoid* is a monoid equipped with an order $\leqslant$ compatible with the multiplication: $x \leqslant y$ implies $zx \leqslant zy$ and $xz \leqslant yz$. Morphisms between ordered monoids are required to be order preserving. A *variety of finite ordered monoids* is a class of finite ordered monoids closed under taking ordered submonoids, quotients and finite direct products.

Reiterman's theorem was adapted to ordered structures in [30]. In particular a class of finite ordered monoids is a variety if and only if it can be defined by a set of profinite identities of the form $u \leqslant v$.

Before stating the extension of Eilenberg's theorem, we need to introduce the syntactic ordered monoid of a language. Let L be a language of A^*. This

notion was first introduced by Schützenberger in 1956 [35], but thereafter, he apparently only used the syntactic monoid.

The *syntactic preorder* of L is the relation $\leqslant_L$ defined on A^* by $u \leqslant_L v$ if and only if, for every $x, y \in A^*$,

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

In earlier papers [5, 8, 11, 12, 17, 18, 23, 25, 26, 27, 28], I used the opposite order, but this one is considered more intuitive by many people. There are also strong topological arguments, which cannot be detailed in this paper, in favour of this order.

Note that the *syntactic congruence* of L is the associated equivalence relation $\sim_L$, defined by $u \sim_L v$ if and only if $u \leqslant_L v$ and $v \leqslant_L u$.

The *syntactic monoid* of L is the quotient $M(L)$ of A^* by $\sim_L$ and the natural morphism $\eta : A^* \to A^*/\sim_L$ is called the *syntactic morphism* of L. The syntactic preorder $\leqslant_L$ induces an order on the quotient monoid $M(L)$. The resulting ordered monoid is called the *syntactic ordered monoid* of L.

For instance, let L be the language $\{a, aba\}$, its syntactic monoid is the monoid $M = \{1, a, b, ab, ba, aba, 0\}$ presented by the relations $a^2 = b^2 = bab = 0$. Its syntactic order is $0 < ab < 1$, $0 < ba < 1$, $0 < aba < a$, $0 < b$.

A *positive variety of languages* is a class of languages $\mathcal{V}$ such that,

(1) for each alphabet A, $\mathcal{V}(A^*)$ is a lattice of languages closed under quotients,

(2) $\mathcal{V}$ is closed under inverses of morphisms.

We can now state the ordered counterpart to Eilenberg's theorem [25].

Theorem 3.1 *Let* $\mathbf{V}$ *be a variety of 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 ordered monoids and positive varieties of languages.*

The conjunction of this theorem and the ordered version of Reiterman's theorem gives the following corollary.

Corollary 3.2 *Any positive variety of languages can be defined by a set of profinite identities of the form* $u \leqslant v$*.*

For instance, the positive variety of regular open sets of the pro-group topology is defined by the identity $1 \leqslant x^\omega$ [24, 26].

Here is another example from [25]. A *shuffle ideal* is a finite union of simple languages. Shuffle ideals form a positive variety, defined by the identity $1 \leqslant x$.

Corollary 3.2 also allows one to study new logic fragments. For instance, shuffle ideals are captured by $\mathbf{\Sigma}_1[<]$, and thus $\mathbf{\Sigma}_1[<] = [\![1 \leqslant x]\!]$.

4 $\mathcal{C}$-varieties

Eilenberg's variety theorem and its ordered version proved to be a powerful tool for studying classes of regular languages. However, some important families of languages arising in language theory do not form positive varieties. To study

some of these classes of languages, Straubing [39] and independently Ésik [14] introduced a more general definition.

Let us say that a morphism $\varphi : A^* \to B^*$ is *length-preserving* (p) if, for each $u \in A^*$, the words u and $\varphi(u)$ have the same length, or equivalently, if for each letter $a \in A$, $\varphi(a)$ is a letter. Similarly, φ is *length-decreasing* (d) if the image of each letter is either a letter or the empty word, and *length-increasing* (i) if the image of each letter is a nonempty word. Finally, it is *length-multiplying* (m) if there is an integer k such that the image of any letter is a word of length k.

Let $\mathcal{C}$ be a class of morphisms between free monoids, closed under composition and containing all p-morphisms. Examples include the classes of all [p-, d-, i-, m-] morphisms.

A [*positive*] $\mathcal{C}$*-variety of languages* is a class $\mathcal{V}$ of regular languages satisfying the first condition defining a [positive] variety but only a weaker version of the second condition of this definition:

($2'$) $\mathcal{V}$ is closed under inverses of $\mathcal{C}$-morphisms.

When $\mathcal{C}$ is the class of p- [d-, i-, m-]morphisms, we use the term p-variety [d-, i-, m-variety].

To obtain a variety theorem for $\mathcal{C}$-varieties, one needs to consider a new type of objects, called stamps. A *stamp* is a surjective morphism from A^* onto a finite monoid. In particular, the *syntactic stamp* of a regular language of A^* is its syntactic morphism. Then one can define an appropriate notion of variety of stamps, which corresponds bijectively to $\mathcal{C}$-varieties [39].

Theorem 4.1 *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.*

Reiterman's theorem can also be extended to this setting [19, 28]. First observe that every morphism from A^* to B^* is uniformly continuous for the profinite metric and thus extends uniquely to a continuous morphism $\widehat{f} : \widehat{A^*} \to \widehat{B^*}$.

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

In practice, one may consider a $\mathcal{C}$-identity as an equation in which each letter represents a variable. If $\mathcal{C}$ is the class of l-morphisms, these variables can be replaced by letters, if it is the class of m-morphisms, they can be replaced by words of the same fixed length, etc., according to the following table:

Class of morphisms	Identity type	Interpretation of variables
all morphisms	identity	words
length preserving	p-identity	words of length 1
length increasing	i-identity	words of length $\geqslant 1$
length decreasing	d-identity	words of length $\leqslant 1$
length multiplying	m-identity	words of the same length

The difference between these different types of identities is illustrated in the following example, taken from [5]. Let L be a regular language over A and let n be its exponent[4]. Then L satisfies the identity [p-, d-, i-, m-identity] $(x^\omega y^\omega)^{\omega+1} = (x^\omega y^\omega)^\omega$ if and only if, for every pair of words [letters, words of length $\leqslant 1$, nonempty words, words of the same length] $(u, v) \in A^* \times A^*$, one has $(u^n v^n)^{n+1} \sim_L (u^n v^n)^n$.

Again we have the following corollary.

Corollary 4.2 *Any [positive] $\mathcal{C}$-variety of languages can be defined by a set of profinite $\mathcal{C}$-identities of the form $u = v$ [$u \leqslant v$].*

Let us give two instances of this theorem. The first example is taken from [12]. Let us call *modular simple* a language of the form

$$(A^d)^* a_1 (A^d)^* a_2 (A^d)^* \cdots a_k (A^d)^*$$

where $d > 0$, $k \geqslant 0$ and $a_1, a_2, \ldots, a_k \in A$. Finite union of modular simple languages form a positive m-variety, defined by the m-identities $1 \leqslant x^{\omega-1}y$ and $1 \leqslant yx^{\omega-1}$.

The second example is related to Boolean circuits. Recall that AC^0 is the set of unbounded fan-in, polynomial size, constant-depth Boolean circuits. Straubing [4, 39] has shown that the regular languages recognised by a circuit in AC^0 form a m-variety defined by the m-identity $(x^{\omega-1}y)^\omega = (x^{\omega-1}y)^{\omega+1}$ [19, 28].

Both examples have a nice interpretation in logic, but we need to enrich our vocabulary by introducing some new predicate symbols, called the modular predicate symbols. The unary symbol $\textsc{mod}_r^d$ assigns to each n the set $\{i < n \mid i \equiv r \bmod d\}$ and the 0-ary symbol $\textsc{length}_r^d$ assigns TRUE to n if $n \equiv r \bmod d$ and FALSE otherwise. We denote by $\mathbf{FO}[< + \mathrm{MOD}]$ the logic obtained by adjoining these modular predicate symbols.

It is known [38] that $\mathbf{FO}[< + \mathrm{MOD}]$ captures exactly the regular languages of AC^0 and that $\mathbf{\Sigma}_1[< + \mathrm{MOD}]$ captures the finite unions of modular simple languages [12]. Therefore these two logic fragments can be defined by the m-identities above-mentioned.

Let us conclude this section with a famous example. *Generalized regular expressions* are defined in the same way as regular expressions, except that complementation is allowed. The *generalized star-height* of a language is the minimal number of nested stars required in an extended regular expression representing the language. It is shown in [29] that the languages of generalized

[4]Recall that the *exponent* of a regular language L of A^* is the smallest integer n such that, for all $u \in A^*$, $u^n \sim_L u^{2n}$.

star-height $\leqslant n$ form a d-variety. Unfortunately, no set of identities defining these d-varieties is known, except for $n = 0$ (see Theorem 2.1). But finding a d-identity satisfied by all languages of generalized star-height $\leqslant 1$ might be a way to solve the long-standing open problem of the existence of a language of generalized star-height 2.

For further developments about $\mathcal{C}$-varieties, we refer the reader to the papers [15, 16, 10, 11, 28].

5 Equations for lattices of regular languages

Although $\mathcal{C}$-varieties were a big progress, they still did not suffice to study certain lattices of regular languages. Pippenger [32] proposed to relax another condition by introducing strains of languages, which share the same properties as varieties of languages except for the closure under quotients by words, which is not required. However, he gave no equational theory for these classes of languages.

In 2008, Gehrke, Grigorieff and the author went a further step in [17]. The remainder of this section is a survey of the results of this paper.

Let L be a regular language of A^* and let $\overline{L}$ be its closure in the free profinite monoid $\widehat{A^*}$. We say that L *satisfies the profinite equation* $u \to v$ if the condition $u \in \overline{L}$ implies $v \in \overline{L}$. It is equivalent to state that $\widehat{\eta}(u) \in \eta(L)$ implies $\widehat{\eta}(v) \in \eta(L)$, where η denotes the syntactic morphism of L. The main result of [17] can be stated as follows.

Theorem 5.1 *A set of regular languages of A^* is a lattice of languages if and only if it can be defined by a set of equations of the form $u \to v$, where u and v are profinite words of $\widehat{A^*}$.*

Writing $u \leftrightarrow v$ for ($u \to v$ and $v \to u$), we get an equational description of the Boolean algebras of languages.

Corollary 5.2 *A set of regular languages of A^* is a Boolean algebra of languages if and only if it can be defined by a set of equations of the form $u \leftrightarrow v$, where $u, v \in \widehat{A^*}$.*

We say that L *satisfies the equation* $u \leqslant v$ if it satisfies the equation $xuy \to xvy$ for all $x, y \in A^*$. We also say that L *satisfies the equation* $u = v$ if it satisfies the equations $u \leqslant v$ and $v \leqslant u$. There is an equivalent definition using the syntactic [ordered] monoid: L satisfies the equation $u = v$ [$u \leqslant v$] if and only if $\widehat{\eta}(u) = \widehat{\eta}(v)$ [$\widehat{\eta}(u) \leqslant \widehat{\eta}(v)$]. This type of equations is adapted to lattices closed under quotients.

Corollary 5.3 *A set of regular languages of A^* is a lattice* [*Boolean algebra*] *closed under quotients if and only if it can be defined by a set of equations of the form $u \leqslant v$ [$u = v$], where $u, v \in \widehat{A^*}$.*

Note that Theorem 5.1 and Corollary 5.3 subsume the Eilenberg-Reiterman theorem and its extension to $\mathcal{C}$-varieties. Indeed, a class of languages is closed under inverses of $\mathcal{C}$-morphisms if and only if its equations are closed under $\mathcal{C}$-morphisms.

Let us give some examples taken from [17]. A *language with zero* is a language whose syntactic monoid has a zero. Regular languages with zero are closed under Boolean operations and residuals, but we need to introduce a new profinite word to get an explicit equational definition of this class.

Let us fix a total order on the alphabet A. Let $u_0, u_1, \ldots$ be the ordered sequence of all words of A^+ in the induced shortlex order. For instance, if $A = \{a, b\}$ with $a < b$, the first elements of this sequence would be

$$1, a, b, aa, ab, ba, bb, aaa, aab, aba, abb, baa, bab, bba, bbb, aaaa, \ldots$$

It is proved in [33, 2] that the sequence of words $(v_n)_{n\geqslant 0}$ defined by

$$v_0 = u_0,\ v_{n+1} = (v_n u_{n+1} v_n)^{(n+1)!}$$

converges to an idempotent ρ_A of the minimal ideal of $\widehat{A^*}$. We can now state:

Proposition 5.4 *A regular language has a zero if and only if it satisfies the equation $x\rho_A = \rho_A = \rho_A x$ for all $x \in A^*$.*

In the sequel, we shall use freely the symbol 0 in equations to mean that a monoid has a zero.

Here is a second example. A language L of A^* is *nondense* if there exists a word $u \in A^*$ such that $L \cap A^* u A^* = \emptyset$. Note that the *full language* A^* is dense and thus nondense languages do not form a lattice. However, regular languages that are either nondense or full do form a lattice closed under quotients and thus Corollary 5.3 applies to this class. One can show that a regular language of A^* is nondense or full if and only if it satisfies the profinite equations $0 \leqslant x$ for all $x \in A^*$. According to our convention, this writing is a shortcut which should be formally replaced by the three equations $x\rho_A = \rho_A = \rho_A x$ and $\rho_A \leqslant x$.

Other examples of interest include the regular *slender languages* (finite union of languages of the form xu^*y, where $x, u, y \in A^*$) and the regular *sparse languages* (finite union of languages of the form $u_0 v_1^* u_1 \cdots v_n^* u_n$, where u_0, v_1, …, v_n, u_n are words) and their complements, the *coslender* [*cosparse*] languages. All these languages have a zero and thus satisfies the profinite equations given in Proposition 5.4. The other profinite equations characterizing these languages are given in the next proposition, in which $i(u)$ denotes the first letter (or *initial*) of a word u.

Proposition 5.5 *Suppose that $|A| \geqslant 2$ and let L be a regular language with zero of A^*.*

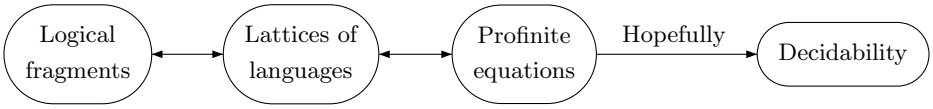
(1) *L is slender or full if and only if it satisfies the equations $0 \leqslant x$ for all $x \in A^*$ and the equation $x^\omega u y^\omega = 0$ for each $x, y \in A^+$, $u \in A^*$ such that $i(uy) \neq i(x)$.*

(2) *L is slender or coslender if and only if it satisfies the equations $x^\omega u y^\omega = 0$ for each $x, y \in A^+$, $u \in A^*$ such that $i(uy) \neq i(x)$.*

(3) *L is sparse or full if and only if it satisfies the equations $0 \leqslant x$ for all $x \in A^*$ and the equations $(x^\omega y^\omega)^\omega = 0$ for each $x, y \in A^+$ such that $i(x) \neq i(y)$.*

(4) *L sparse or cosparse if and only if it satisfies the equations $(x^\omega y^\omega)^\omega = 0$ for each $x, y \in A^+$ such that $i(x) \neq i(y)$.*

Theorem 5.1 also shows that any class of regular languages defined by a fragment of logic closed under conjunctions and disjunctions (first-order, monadic second-order, temporal, etc.) admits an equational description. This opens the way to an algebraic approach to decidability problems in logic, depicted in the following diagram.



To conclude this section, let us mention a recent result of Branco and the author [8], which generalizes earlier results of Weil and the author [31]. Let $\mathcal{L}$ be a lattice of regular languages closed under quotients. The *polynomial closure* of $\mathcal{L}$ is the set of languages that are finite unions of products of the form $L_0a_1L_1\cdots a_nL_n$, where $a_1,\ldots,a_n$ are letters and $L_0,\ldots,L_n$ are languages of $\mathcal{L}$. One can show that it is also a lattice of languages closed under quotient. The next theorem gives a set of equations defining this lattice, given the equations of $\mathcal{L}$.

Theorem 5.6 *The polynomial closure of $\mathcal{L}$ is defined by the set of equations of the form $x^\omega \leqslant x^\omega y x^\omega$, where x, y are profinite words such that the equations $x = x^2$ and $x \leqslant y$ are satisfied by $\mathcal{L}$.*

It would be interesting to get similar results for other closure operators on languages.

6 The final touch

All the results of the previous sections deal with regular languages. Yet it is possible to extend these results to any lattice of languages. The price to pay is to go from profinite equations to another type of equations, the *procompact* equations.

The key space is βA^*, the *Stone-Čech compactification* of A^*. There are many equivalent definitions of this object, and we will remind the reader of three of them. First, one can define βA^* as the set of ultrafilters on the discrete space A^*. A second way is to take the closure of the image of A^* in the product space $\prod K$ where the product runs over all maps from A^* into a compact Hausdorff space K whose underlying set is $\mathcal{P}(\mathcal{P}(A^*))$. That is the reason why we propose to call the elements of βA^* *procompact words*, by analogy with profinite words.

The third construction is a variation of the second one. Let C be the set of all continuous functions from A^* into $[0,1]$. Then the set $[0,1]^C$, equipped with the product topology, is a compact space. Now, the map from A^* into $[0,1]^C$ which maps a word u onto the family $(f(u))_{f\in C}$ is continuous and its image is a compact space, which happens to be precisely βA^*.

This latter construction is reminiscent of the following construction of the profinite monoid $\widehat{A^*}$. Let χ_X be the characteristic function of a subset X of A^*. The set of all functions from A^* into $\{0,1\}$ can be identified with the set $\mathcal{P}(A^*)$. Now, the map from A^* into $\{0,1\}^{\mathcal{P}(A^*)}$ which maps a word u onto the family $(\chi_X(u))_{X\in\mathcal{P}(A^*)}$ is continuous and its image is $\widehat{A^*}$.

Both spaces $\widehat{A^*}$ and βA^* are compact, but only $\widehat{A^*}$ is a compact monoid. However, it is possible to define two continuous maps $\beta A^* \times A^* \to \beta A^*$ and $A^* \times \beta A^* \to \beta A^*$ which both extend the product on A^*. In other words, right and left products by an element of A^* are well defined on βA^*.

Let L be a language of A^* and let $\overline{L}$ be its closure in βA^*. Let u and v be two procompact words of βA^*. We say that L *satisfies the equation* $u \to v$ if $u \in \overline{L}$ implies $v \in \overline{L}$. The following characterization was given in [18]. It is very general since it applies to any lattice of languages.

Theorem 6.1 *A set of languages of A^* is a lattice of languages if and only if it can be defined by a set of equations of the form $u \to v$, where u and v are elements of βA^*.*

For instance, the lattice of right ideals of A^* (languages of the form LA^*) is defined by the set of equations $\{x \to xy \mid x \in \beta A^*, y \in A^*\}$.

7 Conclusion

Equational descriptions open the way to an algebraic approach to some difficult problems. These problems arise in language theory, like the decision problem for a given class of languages, but may also deal with the expressive power of certain logic fragments. However, there are several difficulties to overcome, even for regular languages.

First, Theorem 5.1 guarantees the existence of equations, but does not provide any method for finding them. Second, profinite words might be difficult to handle, in particular when they are only given as the limit of a sequence of words. Third, the set of profinite equations is in general infinite, and even if it is given by some recursive description, one still needs to find an algorithm to decide whether a given regular language satisfies these equations. Nevertheless, these problems have been already partially solved for varieties of languages, where the same difficulties are present, and so one can be reasonably optimistic.

In the nonregular case, Theorem 6.1 shows that one can, in principle, separate any two distinct lattices of languages by exhibiting an equation holding in one and not in the other. The difficulty arises from the problem of getting one's hands on such an equation. Again, the existence statement provided by the theory does not include a means of finding the equations. A further difficulty is that procompact words cannot be given explicitly.

Finally, the scope of equations is not limited to languages of finite words. Actually, Eilenberg-Reiterman theory has already been successfully extended to infinite words and to ω-regular languages [23]. Extensions to words over linear orders [9] and to finite and infinite trees [7] are on the way.

References

[1] J. Almeida, *Finite semigroups and universal algebra*, World Scientific Publishing Co. Inc., River Edge, NJ, 1994. Translated from the 1992 Portuguese original and revised by the author.

[2] J. Almeida and M. V. Volkov, Profinite identities for finite semigroups whose subgroups belong to a given pseudovariety, *J. Algebra Appl.* **2**,2 (2003), 137–163.

[3] B. Banaschewski, The Birkhoff theorem for varieties of finite algebras, *Algebra Universalis* **17**,3 (1983), 360–368.

[4] D. A. M. Barrington, K. Compton, H. Straubing and D. Thérien, Regular languages in NC^1, *J. Comput. System Sci.* **44**,3 (1992), 478–499.

[5] J. Berstel, L. Boasson, O. Carton, J.-É. Pin and A. Restivo, The expressive power of the shuffle product, *Information and Computation* **208** (2010), 1258–1272.

[6] G. Birkhoff, On the structure of abstract algebras., *Proc. Camb. Philos. Soc.* **31** (1935), 433–454.

[7] M. Bojańczyk, L. Segoufin and H. Straubing, Piecewise testable tree languages, in *LICS*, pp. 442–451, IEEE Computer Society, 2008.

[8] M. J. Branco and J.-É. Pin, Equations for the polynomial closure, in *ICALP 2009, Part II*, S. Albers, A. Marchetti-Spaccamela, Y. Matias, S. Nikoletseas and W. Thomas (eds.), Berlin, 2009, pp. 115–126, *Lect. Notes Comp. Sci.* vol. 5556, Springer.

[9] V. Bruyère and O. Carton, Automata on linear orderings, *J. Comput. System Sci.* **73**,1 (2007), 1–24.

[10] 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.

[11] L. Chaubard, J.-É. Pin and H. Straubing, Actions, Wreath Products of $\mathcal{C}$-varieties and Concatenation Product, *Theoret. Comput. Sci.* **356** (2006), 73–89.

[12] L. Chaubard, J.-É. Pin and H. Straubing, First order formulas with modular predicates, in *21st Annual IEEE Symposium on Logic in Computer Science (LICS 2006)*, pp. 211–220, IEEE, 2006.

[13] S. Eilenberg, *Automata, Languages and Machines*, vol. B, Academic Press, New York, 1976.

[14] Z. Ésik, Extended temporal logic on finite words and wreath products of monoids with distinguished generators, in *DLT 2002, Kyoto, Japan*, Ito, Masami et al. (ed.), Berlin, 2002, pp. 43–58, *Lect. Notes Comp. Sci.* n° 2450, Springer.

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

[16] Z. Ésik and K. G. Larsen, Regular languages definable by Lindström quantifiers, *Theoret. Informatics Appl.* **37**,3 (2003), 179–241.

[17] M. Gehrke, S. Grigorieff and J.-É. Pin, Duality and equational theory of regular languages, in *ICALP 2008, Part II*, L. Aceto and al. (eds.), Berlin, 2008, pp. 246–257, *Lect. Notes Comp. Sci.* vol. 5126, Springer.

[18] M. Gehrke, S. Grigorieff and J.-É. Pin, Duality and equational theory of regular languages, in *ICALP 2010, Part II*, S. e. a. Abramsky (ed.), Berlin, 2010, pp. 151–162, *Lect. Notes Comp. Sci.* vol. 6199, Springer.

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

[20] 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.

[21] P. Perkins, Bases for equational theories of semigroups, *J. Algebra* **11** (1969), 298–314.

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

[23] D. Perrin and J.-É. Pin, *Infinite Words*, *Pure and Applied Mathematics* vol. 141, Elsevier, 2004. ISBN 0-12-532111-2.

[24] J.-É. Pin, Polynomial closure of group languages and open sets of the Hall topology, in *21th ICALP*, Berlin, 1994, pp. 424–435, *Lect. Notes Comp. Sci.* n° 820, Springer.

[25] J.-É. Pin, A variety theorem without complementation, *Russian Mathematics (Iz. VUZ)* **39** (1995), 80–90.

[26] J.-É. Pin, Polynomial closure of group languages and open sets of the Hall topology, *Theoret. Comput. Sci.* **169** (1996), 185–200.

[27] J.-É. Pin, Profinite methods in automata theory, in *26th International Symposium on Theoretical Aspects of Computer Science (STACS 2009)*, S. Albers and J.-Y. Marion (eds.), pp. 31–50, Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, 2009.

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

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

[30] J.-É. Pin and P. Weil, A Reiterman theorem for pseudovarieties of finite first-order structures, *Algebra Universalis* **35** (1996), 577–595.

[31] J.-É. Pin and P. Weil, Polynomial closure and unambiguous product, *Theory Comput. Systems* **30** (1997), 383–422. Journal version of the article of ICALP 1995.

[32] N. Pippenger, Regular languages and Stone duality, *Theory Comput. Syst.* **30**,2 (1997), 121–134.

[33] N. R. Reilly and S. Zhang, Decomposition of the lattice of pseudovarieties of finite semigroups induced by bands, *Algebra Universalis* **44**,3-4 (2000), 217–239.

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

[35] M.-P. Schützenberger, Une théorie algébrique du codage, in *Séminaire Dubreil-Pisot, année 1955-56, Exposé No. 15, 27 février 1956, 24 pages*, Inst. H. Poincaré, Paris, 1956. `http://igm.univ-mlv.fr/~berstel/Mps/Travaux/A/1956CodageSemDubreil.pdf`.

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

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

[38] H. Straubing, *Finite automata, formal logic, and circuit complexity*, *Progress in Theoretical Computer Science*, Birkhäuser Boston Inc., Boston, MA, 1994.

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

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