Information and Computation ••• (••••) •••–•••

Contents lists available at ScienceDirect

Information and Computation

www.elsevier.com/locate/yinco

Classes of languages generated by the Kleene star of a word ☆

Laure Daviaud [a,*], Charles Paperman [b,*]

[a] DIMAP, Department of Computer Science, University of Warwick, Coventry CV4 7AL, United Kingdom
[b] Links Team of Inria, Université de Lille, avenue Halley 59650 Villeneuve d'Ascq, France

ARTICLE INFO

Article history:
Received 8 December 2015
Received in revised form 26 September 2017
Available online xxxx

Keywords:
Automata theory
Regular languages
Profinite equations
Kleene star
Decidability

ABSTRACT

In this paper, we study the lattice and the Boolean algebra, possibly closed under quotient, generated by the languages of the form u^*, where u is a word. We provide effective equational characterisations of these classes, i.e. one can decide using our descriptions whether a given regular language belongs or not to each of them.

1. Introduction

The use of equational descriptions of regular languages is a successful and long-standing approach to obtain characterisations of classes of regular languages. One of the first results, following Schützenberger's theorem [14], is the characterization of star-free languages by ultimately equational descriptions given by McNaughton and Papert [8, chapter 4] and later by Eilenberg and Schützenberger [4]. In the case of a variety of regular languages, Reiterman's theorem [13] guarantees the existence of a characteristic set of profinite equations. This theorem has been extended to several kinds of classes of languages, including lattices and Boolean algebras. The reader could refer to [5,10] for a more detailed presentation. Let $\mathcal{U}$ be the class of all languages of the form u^*, where u is a word. The aim of this paper is to study the four classes of regular languages $\mathcal{L}$, $\mathcal{B}$, $\mathcal{L}q$ and $\mathcal{B}q$ obtained respectively as the closure of $\mathcal{U}$ under the following operations: finite union and finite intersection (lattice operations) for $\mathcal{L}$, finite union, finite intersection and complement (Boolean operations) for $\mathcal{B}$, lattice operations and quotients for $\mathcal{L}q$ and Boolean operations and quotients for $\mathcal{B}q$.

Our main result is an equational characterisation for each of these four classes. These equational characterisations being effective, they give as a counterpart the decidability of the membership problem: One can decide whether a given regular language belongs to $\mathcal{L}$, $\mathcal{B}$, $\mathcal{L}q$ and $\mathcal{B}q$ respectively. In addition to describing $\mathcal{L}$, $\mathcal{B}$, $\mathcal{L}q$ and $\mathcal{B}q$ in terms of equations, our results also provide a general form for the languages belonging to each of these classes.

Motivations. Our motivation for the study of these classes is threefold. First, a few years ago, Restivo proposed the problem of characterising the variety of languages generated by the languages of the form u^*, where u is a word.[1] Given that a

☆ This work was carried out when the first author was supported by ANR Project ELICA ANR-14-CE25-0005 and by ANR Project RECRE ANR-11-BS02-0010 (ENS Lyon, France) and the second author by Warsaw Center of Mathematics and Computer Science (WCMCS) (Poland).

* Corresponding authors.
E-mail addresses: l.daviaud@warwick.ac.uk (L. Daviaud), charles.paperman@inria.fr (C. Paperman).

[1] Personal communication to J-É. Pin.

https://doi.org/10.1016/j.ic.2018.07.002
0890-5401/

variety of languages is a class of regular languages closed under Boolean operations, quotients and inverses of morphisms, our result can be viewed as a first step towards the solution of Restivo's problem.

Our second reason for studying these classes was to provide non trivial applications of the equational theory of regular languages as defined by Gehrke, Grigorieff and Pin in [5,10]. There are indeed plenty of examples of known equational characterisations of varieties of languages, but not so many of classes of languages that are not closed under inverses of morphisms or under quotients. As far as we know, the only other studied examples are the ones given in [5] (about languages with a zero and slender languages – see related work below).

Our third motivation is rather a long term perspective since it has to do with the (generalised) star-height problem, a long standing open problem on regular languages [11]. It appears that a key step towards the solution of this problem would be to characterise the Boolean algebra generated by the languages of the form F^*, where F is a finite language. The case $F = \{u\}$ studied in this paper is certainly a very special case, but it gives an insight into the difficulty of the general problem.

Related work. A related class is the class of *slender languages* [6,15], which can be written as finite unions of languages of the form xu^*y, where $x, u, y \in A^*$. The class of *slender or full languages* is a lattice closed under quotients that is therefore characterised by a set of equations [5]. These equations correspond in fact to patterns that cannot be found in any minimal automaton that computes a slender language. In our case, equations provided to characterise classes $\mathcal{L}$, $\mathcal{B}$, $\mathcal{L}q$ and $\mathcal{B}q$ can also be seen as forbidden patterns in automata. Then, we deduce normal forms for the languages in $\mathcal{L}$, $\mathcal{B}$, $\mathcal{L}q$ and $\mathcal{B}q$.

Organisation of the paper. Section 2 gives classical definitions and properties about algebraic automata theory and profinite semigroups. Section 3 is dedicated to the study of the syntactic monoid of u^* for a given word u. In particular, we present useful algebraic properties of the syntactic monoid of u^*. Section 4 presents the equational theory of regular languages: it first gives classical results, then presents the equations satisfied by u^*, and finally gives the characterisations of $\mathcal{L}$, $\mathcal{L}q$ and $\mathcal{B}q$. The study of $\mathcal{B}$ is much more intricate and involves specific tools that are given in Section 5. Finally, Section 6 presents decidability issues. Sections 2 to 6 deal with alphabet with at least two letters. The case of a unary alphabet derives from the two-letter case. It is treated in Section 7.

Notations. We denote by A a finite alphabet with at least two letters, by A^* the set of words on A, by 1 the empty word and by $|u|$ the length of a word u.

This paper is a long version of the paper [3], that was published in MFCS'15, containing all the proofs that were missing in the short version and a complete description of the unary case.

2. Recognisability and the profinite monoid

In this section, we introduce the definitions of recognisability by monoids and of profinite monoid. For more details, the reader could refer to [2].

Monoids and recognisability. A monoid M is a set equipped with a binary associative operation with a neutral element denoted by 1. The product of x and y is denoted by xy. An element e of M is idempotent if $e^2 = e$. An element $0 \in M$ is a zero of M if for all $x \in M$, $0x = x0 = 0$. Given two monoids M and N, $\varphi : M \to N$ is a morphism if for all $x, y \in M$, $\varphi(xy) = \varphi(x)\varphi(y)$ and $\varphi(1) = 1$.

In a finite monoid, every element has an idempotent power: for all $x \in M$, there is $n_x \in \mathbb{N} - \{0\}$ such that x^{n_x} is idempotent. The smallest n_x satisfying this property is called the *index* of x. Moreover, there is an integer $n \neq 0$ such that for all $x \in M$, x^n is idempotent. For instance, one could take the product of the n_x. The smallest integer satisfying this property is called the *index* of the monoid and is denoted by ω. Thus, x^ω is the unique idempotent in the subsemigroup generated by x.

Given a monoid M and a morphism $\varphi : A^* \to M$, a language L is said to be *recognised* by (M, φ) if there is $P \subseteq M$ such that $L = \varphi^{-1}(P)$. The language L is said to be recognised by M if there is φ such that (M, φ) recognises L. A language is regular if and only if it is recognised by a finite monoid. Moreover, the smallest monoid that recognises a regular language L is unique up to isomorphism and is called the *syntactic monoid* of L. The associated morphism φ is called the *syntactic morphism* and $\varphi(L)$ is called the *syntactic image* of L. Furthermore, for each word u, we call $\varphi(u)$ the syntactic image of u with respect to L. The syntactic monoid of a regular language can be computed as it is the transition monoid of the minimal (deterministic) automaton of L.

Free profinite monoid. Given two words u and v, a monoid M *separates* u and v if there is a morphism $\varphi : A^* \to M$ such that $\varphi(u) \neq \varphi(v)$. If $u \neq v$, there is a finite monoid that separates u and v. A distance d can be defined on A^* as follows: $d(u, u) = 0$ and if $u \neq v$, $d(u, v) = 2^{-n}$ where n is the smallest size of a monoid that separates u and v.

Every finite monoid is seen as a metric space equipped with the distance $d(x, y) = 1$ if $x \neq y$ and $d(x, y) = 0$ otherwise. This implies that every morphism from A^* to a finite monoid is a uniformly continuous function.

We briefly recall some useful definitions and results on the *free profinite monoid*. We refer to [2] for an extended presentation of this subject. The free profinite monoid of A^*, denoted by $\widehat{A^*}$ can be defined as the completion for the distance d of A^*. It is a compact space such that A^* is a dense subset of $\widehat{A^*}$. Its elements are called *profinite words*. It is known that a language L is regular if and only if $\overline{L}$ is open and closed in $\widehat{A^*}$, where $\overline{L}$ is the topological closure of L in $\widehat{A^*}$.

Please cite this article in press as: L. Daviaud, C. Paperman, Classes of languages generated by the Kleene star of a word, Inf. Comput. (2018), https://doi.org/10.1016/j.ic.2018.07.002

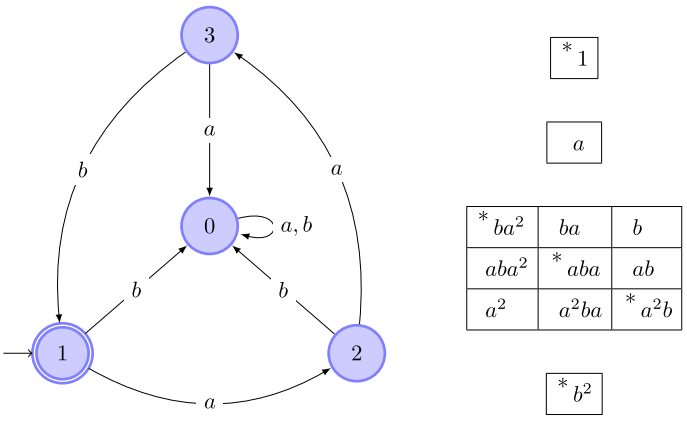

Fig. 1. Minimal deterministic automaton and monoid representation of $(aab)^*$.

Finally, every morphism from A^* to some finite monoid M can be uniquely extended to a uniformly continuous morphism from $\widehat{A^*}$ to M. Given a morphism φ from A^* to M, we will denote by $\widehat{\varphi}$ its unique extension. Moreover, in the rest of the paper, we will use the convention of denoting morphisms from $\widehat{A^*}$ with a $\widehat{\ }$.

The two following examples are profinite words that are not finite words and that will be intensively used in the remainder of the paper.

Example 1 *(Idempotent power).* Given a word $u \in A^*$, the sequence $(u^{n!})_n$ converges in $\widehat{A^*}$. Its limit is denoted by u^ω. Given a finite monoid M and a morphism $\widehat{\varphi} : \widehat{A^*} \to M$, $\widehat{\varphi}(u^\omega) = \widehat{\varphi}(u)^\omega$.

Example 2 *(Zero [1,12]).* Let A be an alphabet with at least two letters and fix a total order on it. Let $(u_n)_n$ be the sequence of all words ordered by the induced shortlex order (a word u is smaller than a word v for the shortlex order if $|u| < |v|$ or $|u| = |v|$ and u is smaller than v for the lexicographic order induced by the order on the alphabet). We set: $v_0 = u_0$ and for all $n \in \mathbb{N}$, $v_{n+1} = (v_n u_{n+1} v_n)^{(n+1)!}$. The sequence $(v_n)_n$ converges in $\widehat{A^*}$ and we denote by ρ_A its limit. Given a finite monoid M and a morphism $\widehat{\varphi} : \widehat{A^*} \to M$, if M has a zero then $\widehat{\varphi}(\rho_A) = 0$.

3. The languages u^*

As mentioned in the introduction, our goal is to describe classes generated by the languages u^*. We will see in Section 4 that proving the correctness of such characterisations requires a precise description of the structure of the syntactic monoid of a given language u^* and particularly of its idempotents.

Therefore, this section addresses this study by exhibiting some properties of the syntactic monoid of u^*. Let us introduce two notions useful to study the languages of this form. A word u is said to be *primitive* if for all words v and all integers n, the condition $u = v^n$ implies $n = 1$ and $v = u$. A word v is said to be a *conjugate* of u if there are words u_1, u_2 such that $u = u_1u_2$ and $v = u_2u_1$.

Every language v^* can thus be written as $(u^m)^*$ where u is a primitive word and m a positive integer. We consider now a primitive word u and a positive integer m.

We will prove that in the syntactic monoid of $(u^m)^*$, there is a zero that is the syntactic image of words that cannot be completed into a word of $(u^m)^*$. Idempotent elements are exactly this zero, the neutral element and the syntactic images of the conjugates of u^m. Thus there are $|u| + 2$ idempotents. Moreover, if the idempotent power of a syntactic image of a word is not zero, then this word has to be a power of a conjugate of u. Finally, the index of the syntactic images of u and of its conjugates is m. All these properties are instantiated in Example 3 and stated in the following propositions.

Example 3. We show in Fig. 1 the minimal deterministic automaton and monoid representation of the language $(aab)^*$. The elements in boxes are the elements of the syntactic monoid of $(aab)^*$. An element has a star in its box if it is idempotent. The conjugates of aab are aab, aba and baa. Finally, the syntactic image of b^2 is a zero of the monoid.

We now introduce some preliminary results about the structure of the syntactic monoid of languages of the form $(u^m)^*$. In what follows, M will denote the syntactic monoid of $(u^m)^*$ and φ its syntactic morphism.

Proposition 1. *The index of u and of its conjugates is m.*

Please cite this article in press as: L. Daviaud, C. Paperman, Classes of languages generated by the Kleene star of a word, Inf. Comput. (2018), https://doi.org/10.1016/j.ic.2018.07.002

Proof. First of all, for $1 \leqslant k < m$ let us prove that $\varphi(u)^k \neq \varphi(u)^{2k}$.

$$u^k u^{m-k} = u^m \in (u^m)^* \quad \text{and} \quad u^{2k} u^{m-k} = u^{m+k} \notin (u^m)^*.$$

Finally let us prove that $\varphi(u)^m = \varphi(u)^{2m}$. Given two words s and t, $su^m t \in (u^m)^*$ if and only if $su^{2m}t \in (u^m)^*$. The same proof holds for the conjugates of u. □

Recall that the alphabet is supposed to contain at least two letters (the following propositions do not hold in the unary case).

Proposition 2. *The monoid M has a zero which is the image of a word that does not belong to $(u^m)^*$.*

Proof. Consider a word v such that $|v| = |u|$ and v is not a factor of u^2. First of all, such a word exists, since u^2 has at most $|u|$ factors of length $|u|$ and there are at least $2^{|u|}$ words of length $|u|$ and $n < 2^n$ for all positive integer n. Finally, $\varphi(v) = 0$. Indeed, for all words s, t we have $svt \notin (u^m)^*$. Otherwise v would be a factor of u^2. □

Proposition 3. *The profinite word ρ_A does not belong to $\overline{(u^m)^*}$.*

Proof. By Proposition 2, the syntactic monoid of $(u^m)^*$ has a zero that is the image of a word that does not belong to $(u^m)^*$. Thus 0 is not in the syntactic image of $(u^m)^*$, and since $\widehat{\varphi}(\rho_A) = 0$ then $\rho_A \notin \overline{(u^m)^*}$. □

The following proposition can be found for example in [7, Proposition 1.3.3].

Proposition 4. *The word u^m has exactly $|u|$ conjugates that are the words of the form v^m where v is a conjugate of u.*

Proposition 5. *Non-zero idempotents in M are either the neutral element of M or the syntactic images of the conjugates of u^m. Thus there are exactly $|u| + 2$ idempotents in the syntactic monoid of $(u^m)^*$.*

Proof. First of all, given a conjugate v of u^m and two words s and t, $svt \in (u^m)^*$ if and only if $sv^2t \in (u^m)^*$. Conversely, consider a word v such that for all s, t, $svt \in (u^m)^*$ if and only if $sv^2t \in (u^m)^*$. Then, if $\varphi(v) \neq 0$ then there is s and t such that both $svt \in (u^m)^*$ and $sv^2t \in (u^m)^*$. Thus there is k such that v is a conjugate of u^{km}, that means that $v = v'^{km}$ where v' is a conjugate of u. By Proposition 1, we have: $\varphi(v) = \varphi(v')^{km} = \varphi(v'^m)$. Thus, a non-zero idempotent in M is the syntactic image of a conjugate of u^m. □

Proposition 6. *For all words v, if $\varphi(v)^\omega \neq 0$ then v is a power of a conjugate of u.*

Proof. Assume that $\varphi(v)^\omega \neq 0$ and let k be an integer such that $\varphi(v)^k = \varphi(v)^\omega$. By Propositions 4 and 5, there exist two words s, t such that $u = st$ and $\varphi(v^k) = \varphi((ts)^m)$. Moreover,

$$u^{2m} = (st)^{2m} = s(ts)^m(ts)^{m-1}t$$

Therefore, there exists k' such that:

$$sv^k(ts)^{m-1}t = (st)^{k'm}$$

and so:

$$v^k = (ts)^{(k'-1)m} \quad \square$$

4. Equational characterisations of $\mathcal{L}$, $\mathcal{L}q$ and $\mathcal{B}q$

This section covers the equational theory of regular languages. First, Section 4.1 presents known results about equations. Then Section 4.2 applies this theory to the study of $\mathcal{L}$, $\mathcal{L}q$ and $\mathcal{B}q$, by giving equations that characterise them.

4.1. Equational characterisations of algebraic structures of regular languages

A *lattice* (resp. a *Boolean algebra*) of languages of A^* is a class of languages containing the empty language $\emptyset$, the full language A^* and which is closed under finite union and finite intersection (resp. finite union, finite intersection and complement). A class of languages $\mathcal{L}$ is *closed under quotients* if for all $L \in \mathcal{L}$, for all $w \in A^*$, $w^{-1}L$ and Lw^{-1} belong to $\mathcal{L}$. Recall that $w^{-1}L = \{w' \mid ww' \in L\}$ and $Lw^{-1} = \{w' \mid w'w \in L\}$. Let u and v be two profinite words. A language $L \subseteq A^*$ satisfies the equation $u \to v$ if the condition $u \in \overline{L}$ implies $v \in \overline{L}$. It satisfies $u \leqslant v$ if for all words x, y, $xuy \in \overline{L}$ implies $xvy \in \overline{L}$. The

notation $u \leftrightarrow v$ is shorthand for $u \to v$ and $v \to u$ and similarly $u = v$ is shorthand for $u \leqslant v$ and $v \leqslant u$. Observe that given a regular language L and its syntactic morphism $\varphi : A^* \to M$, the language L satisfies $u = v$ if and only if $\widehat{\varphi}(u) = \widehat{\varphi}(v)$ in M. A class of languages $\mathcal{L}$ is defined by a set of equations E if the following equivalence holds: $L \in \mathcal{L}$ if and only if L satisfies all the equations in E.

The kind of equations used to describe a class of languages is strongly related to its closure operations. The two following propositions formalise this statement.

Proposition 7 *(Theorem 5.2 [5]). A class of regular languages is defined by a set of equations of the form $u \to v$ (resp. $u \leftrightarrow v$) if and only if it is a lattice (resp. a Boolean algebra) of regular languages.*

Proposition 8 *(Theorem 7.2 [5]). A class of regular languages is defined by a set of equations of the form $u \leqslant v$ (resp. $u = v$) if and only if it is a lattice (resp. a Boolean algebra) of regular languages closed under quotients.*

Equations with zero. The existence of a zero in a syntactic monoid is given by the equations:

$$\rho_A x = x\rho_A = \rho_A$$

If these equations are satisfied, we will use the notation 0 instead of ρ_A. For example the set of equations:

$$\begin{cases} \rho_A \leqslant x \\ \rho_A x = x\rho_A = \rho_A \end{cases} \quad \text{is replaced by } 0 \leqslant x$$

The zero has been used to describe several classes of languages. For instance, the equations $0 \leqslant x$ for $x \in A^*$ describe exactly the so called *nondense* languages. Another example is the class of slender or full languages defined in the introduction [6,15]. This class of languages is a lattice closed under quotients; it is described by the following equations:

$$0 \leqslant x \text{ for } x \in A^*$$
$$x^\omega u y^\omega = 0 \text{ for } x, y \in A^+, u \in A^* \text{ and } i(uy) \neq i(x)$$

where $i(v)$ is the first letter of v for any $v \in A^+$ [9].

4.2. Characterisations of $\mathcal{L}$, $\mathcal{L}q$ and $\mathcal{B}q$

We give here a list of equations used in the study of $\mathcal{L}$, $\mathcal{L}q$, $\mathcal{B}q$ and $\mathcal{B}$. The proofs of the characterisations of these classes by sets of equations are made in two steps.

We first verify that the equations are correct and then check for their completeness. For the first step, it is sufficient to prove that for all words u, the language u^* satisfies the set of equations. From the nature of the equations ($\to$, $\leftrightarrow$, $\leqslant$, $=$), we then obtain directly that the whole lattice, Boolean algebra and their closure under quotients satisfy the given set of equations. This step of correctness can be derived from the structure of the languages of the form u^*, presented in Section 3.

The second step is to prove that only the languages in the desired structures satisfy the set of equations. This step is more intricate since it requires a full understanding of the combinatorics of the classes we consider.

First we define the following two languages:

$$P_u = \bigcup_{p \text{ prefix of } u} u^* p \quad \text{and} \quad S_u = \bigcup_{s \text{ suffix of } u} s u^*$$

The equations:

$$x^\omega y^\omega = 0 \text{ for } x, y \in A^* \text{ such that } xy \neq yx \tag{E_1}$$
$$x^\omega y = 0 \text{ for } x, y \in A^* \text{ such that } y \notin P_x \tag{E_2}$$
$$y x^\omega = 0 \text{ for } x, y \in A^* \text{ such that } y \notin S_x \tag{E_3}$$
$$x^\omega \leqslant 1 \text{ for } x \in A^* \tag{E_4}$$
$$0 \leqslant 1 \tag{E_5}$$
$$x^\ell \leftrightarrow x^{\omega+\ell} \text{ for } x \in A^*,\ \ell > 0 \tag{E_6}$$
$$x^\omega \to 1 \text{ for } x \in A^* \tag{E_7}$$
$$x \to x^\ell \text{ for } x \in A^*,\ \ell > 0 \tag{E_8}$$

Please cite this article in press as: L. Daviaud, C. Paperman, Classes of languages generated by the Kleene star of a word, Inf. Comput. (2018), https://doi.org/10.1016/j.ic.2018.07.002

Some equations are clearly satisfied by u^* such as equations (E_8) and (E_7). Indeed, if $v \in u^*$ then for all ℓ, v^ℓ is also a power of u and belongs to u^* (E_8). Similarly, 1 always belongs to u^* (E_7).

Proving that u^* satisfies the other equations is more difficult and requires to analyse the structure of its syntactic monoid. In particular, the role of the idempotents is important. These proofs are given in the following propositions. In particular, we first give some implications between these sets of equations.

Proposition 9.

1. *The sets of equations* (E_1) *and* (E_6) *imply the set of equations* (E_2).
2. *The sets of equations* (E_1) *and* (E_6) *imply the set of equations* (E_3).
3. *The sets of equations* (E_1) *and* (E_4) *imply the set of equations* (E_5).
4. *The set of equations* (E_4) *implies the set of equations* (E_7).
5. *The sets of equations* (E_4) *and* (E_8) *imply the set of equations* (E_6).

Proof. *1.* Let L be a regular language satisfying the sets of equations (E_1) and (E_6). Let us denote by φ its syntactic morphism, m the index of its syntactic monoid and P its syntactic image. Either $0 \in P$ or not. If $0 \in P$, then we can argue on L^c. Indeed, L^c has the same syntactic monoid and its syntactic image is P^c. Thus $0 \notin P^c$, and since the sets of equations (E_1) and (E_6) are symmetric ($=$ and $\leftrightarrow$), then L^c also satisfies them. Finally, since (E_2) is also symmetric, we have that if L^c satisfies it, then L will also satisfy it. So, we can suppose that $0 \notin P$.

Assume now that $\widehat{\varphi}(x^\omega y) \neq 0$ for some words x and y. Then this implies that there are words u and v such that $ux^\omega yv \in \overline{L}$. By (E_6), $(ux^\omega yv)^{\omega+1} \in \overline{L}$. Moreover:

$$(ux^\omega yv)^{\omega+1} = ux^\omega(yvux^\omega)^\omega yv \in \overline{L}$$

Since $0 \notin P$ then $\widehat{\varphi}(x^\omega(yvux^\omega)^\omega) \neq 0$ or equivalently $\widehat{\varphi}(x^\omega(yvux^m)^\omega) \neq 0$. So by (E_1), $xyvux^m = yvux^m x$. Then, there is a word t and two integers $\ell \leqslant k$ such that $x = t^\ell$ and $yvux^m = t^k$. Finally y is a prefix of t^k so a prefix of $t^{k\ell}$ and so a prefix of x^k. We have proved that if $\widehat{\varphi}(x^\omega y) \neq 0$ then $y \in P_x$ which proves (E_2).

2. The proof is symmetric to the previous one.

3. The set of equations (E_1) is only used to prove the existence of a zero in the monoid. Then it is sufficient to observe that (E_5) is a particular case of equations in (E_4) where x is such that $x^\omega = 0$.

4. For all words x and y, $x \leqslant y$ implies by definition that $x \to y$.

5. Let $\ell > 0$ and $x \in A^*$. The set of equations (E_4) implies that for all words u and v, $ux^\omega v \to uv$. By taking $u = 1$ and $v = x^\ell$, we obtain that $x^{\omega+\ell} \to x^\ell$. Conversely, let L be a regular language satisfying the set of equations (E_8), and m the index of its syntactic monoid. Suppose that $x^\ell \in L$. Then by (E_8), $(x^\ell)^{m+1} = x^{\ell m+\ell} \in L$. Finally this implies that $x^{\omega+\ell} \in \overline{L}$, which proves $x^\ell \to x^{\omega+\ell}$. □

Proposition 10. *For all words* u, *the language* u^* *satisfies the sets of equations* (E_1), (E_2), (E_3), (E_4), (E_6) *and* (E_8).

Proof. (E_1)*:* Let $x, y \in A^*$ such that $xy \neq yx$, and $u = v^m$ with v primitive. Let us denote by φ the syntactic morphism of u^*. First of all, if $\varphi(x)^\omega = 0$ or $\varphi(y)^\omega = 0$, then the equation is satisfied. Otherwise, by Proposition 6, x and y are powers of conjugates of v. But since $xy \neq yx$, these conjugates are different and thus $\varphi(x)^\omega\varphi(y)^\omega = 0$.

(E_4)*:* Let $x \in A^*$ and k its idempotent power. Let $u = v^m$ where v is primitive. First assume that $\widehat{\varphi}(x^\omega) \neq 0$. By Proposition 6, there exist two words s and t such that $st = v$ and an integer k' such that $x^k = (ts)^{k'm}$. Let p and q be two words such that $px^kq \in u^*$. There is an integer ℓ such that:

$$px^kq = p(ts)^{k'm}q = (st)^{\ell m}$$

Thus, there exists a prefix s' of st and an integer α such that $p = (st)^\alpha s'$ and a suffix t' and an integer β such that $q = t'(st)^\beta$. Therefore,

$$px^kq = (st)^\alpha s'(ts)^{k'm}t'(st)^\beta = (st)^{\ell m}$$

Necessarily, $s' = s$ and $t' = t$. Finally,

$$pq = (st)^{\alpha+\beta+1} = (st)^{(\ell-k')m} \in u^*$$

So $px^\omega q \in \overline{u^*}$ implies that $pq \in \overline{u^*}$ and thus $x^\omega \leqslant 1$. Finally, if $\widehat{\varphi}(x^\omega) = 0$ then $x^\omega \leqslant 1$ since $0 \notin \overline{u^*}$.

(E_8)*:* If $x \in u^*$, then $x = u^k$ for some integer k. Then for all $\ell > 0$, $x^\ell = u^{k\ell} \in u^*$.

(E_6)*:* By Proposition 9.5 and by (E_4) and (E_8).

(E_2)*:* By Proposition 9.1 and by (E_1) and (E_6).

(E_3)*:* By Proposition 9.2 and by (E_1) and (E_6). □

Please cite this article in press as: L. Daviaud, C. Paperman, Classes of languages generated by the Kleene star of a word, Inf. Comput. (2018), https://doi.org/10.1016/j.ic.2018.07.002

We will prove now that the languages of the form $(u^m)^*u^r$ satisfy some of the sets of equations given in Section 4.2.

Proposition 11. *Given a word u and two positive integers m and r, languages $(u^m)^*$ and $(u^m)^*u^r$ are mutually quotients of one another. Thus they have the same syntactic monoid and verify the same equations of the form $\leqslant$ and $=$.*

Proof. The following two equations prove that languages $(u^m)^*$ and $(u^m)^*u^r$ are mutually quotients of one another.

$$(u^m)^* = ((u^m)^*u^r)(u^r)^{-1} \quad \text{and} \quad (u^m)^*u^r = (u^m)^*(u^{m-r})^{-1} \quad \square$$

Proposition 12. *For all words u, integers m and r, the language $(u^m)^*u^r$ satisfies the sets of equations (E_5), (E_6) and (E_7).*

Proof. (E_5): By Proposition 10, the language $(u^m)^*$ satisfies (E_1) and (E_4). Thus by Proposition 9.3, it also satisfies (E_5). Finally, by Proposition 11, $(u^m)^*u^r$ satisfies the same equations of the form $\leqslant$ and $=$. So it satisfies (E_5).

(E_6): Let x be a word in A^* and assume that $x^\ell = (u^m)^k u^r$ for $k \in \mathbb{N}$. Since x is a power of u, its idempotent power, x^m, is syntactically equivalent to u^m. Thus:

$$\widehat{\varphi}(x^\omega x^\ell) = \varphi(u^m x^\ell) = \varphi((u^m)^{k+1}u^r) \in \varphi(L)$$

By Proposition 11, the language $(u^m)^*u^r$ satisfies $x^\omega \leqslant 1$ since (E_4) is satisfied by $(u^m)^r$. In particular, $x^\omega x^\ell \to x^\ell$ is satisfied. Finally $x^\omega x^\ell \leftrightarrow x^\ell$ (E_6) is satisfied by the language $(u^m)^*u^r$.

(E_7): By Proposition 10, the language $(u^m)^*$ satisfies (E_4). By Proposition 11, the language $(u^m)^*u^r$ also satisfies (E_4). Finally, by Proposition 9.4 it also satisfies (E_7). $\square$

The following theorem gives the equational characterisations of $\mathcal{B}q$, $\mathcal{L}q$ and $\mathcal{L}$.

Theorem 1. *Over a finite alphabet with at least two letters:*

1. *The class $\mathcal{B}q$ is defined by equations (E_1), (E_2) and (E_3).*
2. *The class $\mathcal{L}q$ is defined by equations (E_1), (E_2), (E_3) and (E_4).*
3. *The class $\mathcal{L}$ is defined by equations (E_1), (E_4) and (E_8).*

To prove these characterisations we introduce a normal form for the languages in $\mathcal{B}q$, $\mathcal{L}q$ and $\mathcal{L}$. More precisely, we prove that a language which satisfies the sets of equations can be written in a normal form. Finally, normal forms imply membership in the classes $\mathcal{B}q$, $\mathcal{L}q$ or $\mathcal{L}$.

Remark 1. The proof is constructive: assuming that a language L satisfies the set of equations, one can compute the words and the integers given in the normal form.

We start with the most general class $\mathcal{B}q$ and then we restrict to the classes $\mathcal{L}q$ and $\mathcal{L}$ by adding sets of equations in the equational characterisation. Hence, let us start with $\mathcal{B}q$.

The case of $\mathcal{B}q$

First, we observe that the finite languages are in $\mathcal{B}q$, as for instance, the language $\{aab\}$. Indeed, $\{aab\} = a^{-1}(aaab)^* \cap (aab)^*$. Given a word u, and a non-negative integer r, we denote by $u^{\geqslant r}$ the language u^*u^r. Since this language can be rewritten as $u^* - \{1, u, \dots, u^{r-1}\}$, it belongs to $\mathcal{B}q$. Similarly, by using the closure by quotient we capture the languages $u^{\geqslant r}p$ and $su^{\geqslant r}$ where p (resp. s) is a prefix (resp. a suffix) of u.

Finally, the following normal form fully characterises the class $\mathcal{B}q$: if L is a language in $\mathcal{B}q$ different from A^*, then L can be written as

$$\left(\bigcup_{i=1}^{k} u_i^{\geqslant r_i} p_i\right) \cup F \text{ or } \left(\left(\bigcup_{i=1}^{k} u_i^{\geqslant r_i} p_i\right) \cup F\right)^c$$

where $(u_i)_{i=1\dots k}$ and F are finite sets of words, p_i is a prefix of u_i and $(r_i)_{i=1\dots k}$ are integers. We have sketched the proof that all the languages that can be written in this normal form are in $\mathcal{B}q$. The difficult part is to prove that every regular language that satisfies the equations can be written in the normal form. Let us now give the complete and formal proof. In the following two propositions, recall that $\mathcal{U}$ denote the set of all the languages of the form u^*, where u is a word.

Proposition 13. *The class $\mathcal{L}q$ and thus $\mathcal{B}q$ contains all finite languages.*

Please cite this article in press as: L. Daviaud, C. Paperman, Classes of languages generated by the Kleene star of a word, Inf. Comput. (2018), https://doi.org/10.1016/j.ic.2018.07.002

Proof. Let u be a word, let us prove that $\{u\}$ is generated by quotient and intersection of languages in $\mathcal{U}$. Let a be a letter distinct from the first letter of u. Then $a^{-1}(au)^* \cap u^* = \{u\}$. Indeed,

$$a^{-1}(au)^* \cap u^* = u(au)^* \cap (\{1, u\} \cup uu^*) = \{u\}$$

since the first letter of u is different from a. Finally finite languages are finite unions of singletons. □

Proposition 14. *For all words u and p a prefix of u different from u, the language u^*p is generated by a quotient of languages in $\mathcal{U}$.*

Proof. Let u be a word and p be a prefix of u. Let s be the word such that $u = ps$. Then $u^*p = u^*s^{-1}$. □

Proposition 15. *For all words u, p a prefix of u and r a non-negative integer, $u^{\geqslant r}p \in \mathcal{B}q$.*

Proof. Let u be a word, p be a prefix of u and r be a non-negative integer. If $p = u$ then $u^{\geqslant r}p = u^{\geqslant r+1}$. Hence we can suppose that p is different from u. The language $\{u^n p \mid 0 \leqslant n < r\}$ is finite and thus by Proposition 13 it belongs to $\mathcal{B}q$. Finally $u^{\geqslant r}p = u^*p - \{u^n p \mid 0 \leqslant n < r\}$ belongs to $\mathcal{B}q$ since $u^*p \in \mathcal{B}q$ thanks to Proposition 14. □

Theorem 2 gives two characterisations of $\mathcal{B}q$, the first one in terms of profinite equations and the second one providing a normal form for the languages in $\mathcal{B}q$. The following proposition is needed in the proof of Theorem 2.

Proposition 16 *(Folklore). Let u be a word. Let M be a monoid and $\varphi : A^* \to M$ be a morphism. If $|u| \geqslant |M| + 1$ then there are three words u_1, u_2, u_3 such that $u = u_1u_2u_3$ and for all $k \in \mathbb{N}$, $\varphi(u) = \varphi(u_1u_2^k u_3)$.*

Theorem 2. *Let L be a regular language and m be the index of its syntactic monoid. The following properties are equivalent:*

1. *$L \in \mathcal{B}q$.*
2. *L satisfies the sets of equations (E_1), (E_2) and (E_3).*
3. *If $L \neq A^*$ then there are:*
   - *a finite set of words $(u_i)_{i=1\ldots k}$,*
   - *for all $i = 1, \ldots, k$, a prefix of u_i denoted by p_i,*
   - *a finite set of non negative integers $(r_i)_{i=1\ldots k}$, such that for all i, $r_i < 2m$,*
   - *a finite set of words F,*

   such that:

$$L = \Big(\bigcup_{i=1}^{k} u_i^{\geqslant r_i} p_i\Big) \cup F \quad \text{or} \quad L = \Big(\Big(\bigcup_{i=1}^{k} u_i^{\geqslant r_i} p_i\Big) \cup F\Big)^c$$

Proof of Theorem 2. We prove Theorem 2 by proving the following sequence of implications $1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 1$.

1 *implies* 2*:* By Proposition 10, u^* satisfies equations (E_1), (E_2) and (E_3).

3 *implies* 1*:* Thanks to Propositions 13 and 15, finite languages and languages of the form $u^{\geqslant r}p$ for p a prefix of u belong to $\mathcal{B}q$. Hence, by finite union and complement, languages of the form given in point 3 of the proposition belong to $\mathcal{B}q$.

2 *implies* 3*:* Let L be a language different from A^* which satisfies (E_1), (E_2) and (E_3). Let φ be its syntactic morphism, m be the index of its syntactic monoid and P be its syntactic image. Without loss of generality, we can assume that $0 \notin P$. Otherwise we can argue on L^c since it satisfies the same symmetric equations. We now set a total order on A and consider the induced shortlex order. We construct sequences of pairs of words $(u_n, p_n)_{n\geqslant 1}$ and integers $(r_n)_{n\geqslant 1}$ satisfying the following conditions: for an integer $i \geqslant 1$, u_i is the smallest word for the shortlex order, p_i is the smallest prefix of u_i and r_i is the smallest integer such that:

- $r_i < 2m$,
- $u_i^{\geqslant r_i} p_i \subseteq L$,
- $u_i^{\geqslant r_i} p_i \nsubseteq \left(\cup_{j<i} u_j^{\geqslant r_j} p_j\right)$.

First observe that for a given u_i, there is a finite set of possible p_i. Moreover, for given u_i and p_i, there is at most one possible r_i. Consider now the language:

$$L' = \Big(\bigcup_{i\geqslant 1} u_i^{\geqslant r_i} p_i\Big) \cup \{u \in L \mid |u| \leqslant N\}$$

where N is the size of the syntactic monoid of L. To conclude the proof we prove now that:

Please cite this article in press as: L. Daviaud, C. Paperman, Classes of languages generated by the Kleene star of a word, Inf. Comput. (2018), https://doi.org/10.1016/j.ic.2018.07.002

1. $L = L'$,
2. and the sequence of (u_i) is finite.

First we deal with the equality $L = L'$.

Lemma 1. $L = L'$

Proof. By construction $L' \subseteq L$. Let us now prove the converse direction. Let $v \in L$. If $|v| \leqslant N$ then $v \in L'$ and we can conclude. Otherwise, by Proposition 16, $v = v_1 v_2 v_3$ such that for all $n \in \mathbb{N}$, $\varphi(v) = \varphi(v_1 v_2^n v_3)$. Since $v \in L$ then for all integer n, $v_1 v_2^n v_3 \in L$ and $v_1 v_2^\omega v_3 \in \overline{L}$. Recall that we assume 0 not to be in $\overline{L}$. Therefore we have $\widehat{\varphi}(v_1 v_2^\omega v_3) \neq 0$ and by equations (E_2) and (E_3) we have $v_1 \in S_{v_2}$ and $v_3 \in P_{v_2}$. We set the following notations:

$$\begin{cases} v_1' \text{ is the prefix of } v_2 \\ v_1'' \text{ is the suffix of } v_2 \end{cases} \quad \text{such that } v_1 = v_1'' v_2^k \text{ for some integer } k \text{ and } v_1' v_1'' = v_2.$$

$$\begin{cases} v_3' \text{ is the suffix of } v_2 \\ v_3'' \text{ is the prefix of } v_2 \end{cases} \quad \text{such that } v_3 = v_2^\ell v_3'' \text{ for some integer } \ell \text{ and } v_3'' v_3' = v_2.$$

Therefore we have:

$$v_1 v_2^n v_3 = v_1'' v_2^k v_2^n v_2^\ell v_3'' = v_1''(v_1' v_1'')^{k+1+\ell+n} v_3'' = (v_1'' v_1')^{k+1+\ell+n} v_1'' v_3''$$

Both words v_1' and v_3'' are prefixes of v_2. If $|v_3''| \leqslant |v_1'|$ then $v_1'' v_3''$ is a prefix of $v_1'' v_1'$. Otherwise, $v_1'' v_3'' = v_1'' v_1' p(v_1'')$ where $p(v_1'')$ is a prefix of v_1'. In any case, for all n, $v_1 v_2^n v_3$ can be written as $u^{r+n} p$ for some word u, p a prefix of u and some integer r. Since for all n, $\varphi(v) = \varphi(v_1 v_2^n v_3)$, then $u^{\geqslant r} p \subseteq L$ (and $v = u^{r+1} p$). Moreover, by definition of the index m, if $r = qm + s$ with $s < m$ then $\varphi(u^{r+n}) = \varphi(u^{m+s+n})$. Thus we can choose $r < 2m$. Finally, either u has been chosen in the sequence (since it satisfies the two first conditions) and thus v belongs to L', or $u^{\geqslant r} p \subseteq (\cup_{j<i} u_j^{\geqslant r_j} p_j)$ for some i and thus v also belongs to L'. We have proved that $L = L'$. □

Let us prove now that the sequence $(u_i)_i$ is finite.

Lemma 2. *The sequence $(u_i)_i$ is finite.*

Proof. By contradiction, suppose that we can construct an infinite sequence of words $(u_i)_i$. Then there is an infinite sub-sequence of $(u_i)_i$, denoted by $(v_i)_i$ such that for all i, j, $\varphi(v_i) = \varphi(v_j)$. Suppose that the sequence of $(v_i)_i$ is maximal, that is to say we consider all the words in the sequence $(u_i)_i$ having the same given image by φ. Recall that for all i, there is a prefix p_i of v_i and an integer r_i such that:

$$(v_i)^{\geqslant r_i} p_i \subseteq L \quad \text{and} \quad (v_i)^{\geqslant r_i} p_i \nsubseteq \bigcup_{j<i} (v_j)^{\geqslant r_j} p_j$$

First of all, if there is $i \neq j$ such that $v_i v_j \neq v_j v_i$ then:

$$\widehat{\varphi}(v_i^\omega) = \varphi(v_i)^\omega \varphi(v_j)^\omega = 0 \qquad \text{(by } (E_1)\text{)}$$

Then, $\widehat{\varphi}(v_i^\omega p_i) = 0$ and $0 \in P$, which contradicts the hypothesis. Hence, for all i, j, $v_i v_j = v_j v_i$. Thus there is word v, and a sequence of positive integers $(k_i)_i$ such that $v_i = v^{k_i}$. Nevertheless, one can choose $k_i \leqslant 2m$, since $\varphi(v^{2m+n}) = \varphi(v^{m+n})$. Thus finally, the sub-sequence $(v_i)_i$ is finite, which contradicts the hypothesis, and concludes the proof. □

Example 4. The language $A^* aa A^*$ is not in $\mathcal{B}q$. Indeed, the first equation is not satisfied since the syntactic image of the words ab and b are idempotents, but the syntactic image of abb is not syntactically equal to 0. However, the language $A^*(aa + bb)A^*$ satisfies the three sets of equations and is therefore in $\mathcal{B}q$ but not in $\mathcal{L}q$ since the set of equations (E_4) is not satisfied: the syntactic image of aa is 0, and by equation (E_4), $0 \leqslant 1$, so 1 should be in the language but that is not the case. We can even give the normal form of this language:

$$A^*(aa + bb)A^* = ((ab)^* \cup (ab)^* a \cup (ba)^* \cup (ba)^* b)^c$$

Please cite this article in press as: L. Daviaud, C. Paperman, Classes of languages generated by the Kleene star of a word, Inf. Comput. (2018), https://doi.org/10.1016/j.ic.2018.07.002

The case of $\mathcal{L}q$

We can now achieve the reduction from $\mathcal{B}q$ to $\mathcal{L}q$, that is removing the closure by complement, by adding the set of equations (E_4) in the equational characterisation. Furthermore, we obtain that the normal form is a restriction of the previous one: if $L \in \mathcal{L}q$ is different from A^*, then

$$L = \left(\bigcup_{i=1}^{k} u_i^* p_i \right) \cup F$$

Theorem 3. *Let L be a regular language. The following properties are equivalent:*

1. $L \in \mathcal{L}q$.
2. *L satisfies the set of equations (E_1), (E_2), (E_3) and (E_4).*
3. *If $L \neq A^*$ then there are:*
   - *a finite set of words $(u_i)_{i=1\ldots k}$,*
   - *for all $i = 1, \ldots, k$, a prefix of u_i denoted by p_i different from u_i,*
   - *a finite set of words F,*

   such that:

$$L = (\bigcup_{i=1}^{k} u_i^* p_i) \cup F$$

Proof. 1 *implies* 2*:* By Proposition 10, u^* satisfies (E_1), (E_2), (E_3) and (E_4).

3 *implies* 1*:* By Proposition 13, finite languages belong to $\mathcal{L}q$ and by Proposition 14, languages of the form u^*p for p a prefix of u different from u, also belong to $\mathcal{L}q$. Finally, since languages given in point 3 are finite union of languages in $\mathcal{L}q$, they also belong to $\mathcal{L}q$.

2 *implies* 3*:* Let L be a language different from A^* satisfying (E_1), (E_2), (E_3) and (E_4). Let us denote by m the index of the syntactic monoid of L. First of all, by Theorem 2, since L satisfies (E_1), (E_2) and (E_3) then there exists:

- a finite set of words $(u_i)_{i=1\ldots k}$,
- for all $i = 1, \ldots, k$, a prefix of u_i denoted by p_i,
- a finite set of integers $(r_i)_{i=1\ldots k}$ such that for all i, $r_i < 2m$,
- a finite set of words F,

such that:

$$L = (\bigcup_{i=1}^{k} u_i^{\geqslant r_i} p_i) \cup F \quad \text{or} \quad L = ((\bigcup_{i=1}^{k} u_i^{\geqslant r_i} p_i) \cup F)^c$$

A particular case of (E_4) is $0 \geqslant 1$, which means that if $0 \in \overline{L}$ then $L = A^*$. Indeed, for all words u, $0u = 0 \in \overline{L} \Rightarrow 1u = u \in \overline{L}$. Furthermore, either $0 \in \overline{L}$ and $L = A^*$ or $0 \notin \overline{L}$. Therefore, we have:

$$L \neq ((\bigcup_{i=1}^{k} u_i^{\geqslant r_i} p_i) \cup F)^c$$

Let $i = 1, \ldots, k$, since $u_i^{\geqslant r_i} p_i \subseteq L$ then for all ℓ, $u_i^{m+\ell} p_i \in L$ and thus $u_i^{\omega+\ell} p_i \in \overline{L}$. By (E_4), this implies that for all ℓ, $u_i^{\ell} p_i \in \overline{L}$. Thus $u_i^* p_i \subseteq L$. Finally, we obtained that:

$$L = (\bigcup_{i=1}^{k} u_i^* p_i) \cup F \quad \square$$

The case of $\widetilde{\mathcal{B}}$ and $\widetilde{\mathcal{L}}$

In order to study $\mathcal{L}$ and $\mathcal{B}$, we have to remove the "closure under quotients" from the characterisations above. We deal with these cases by introducing an intermediate Boolean algebra (resp. lattice) denoted by $\widetilde{\mathcal{B}}$ (resp. $\widetilde{\mathcal{L}}$). The latter classes are generated by the following languages, which correspond to a certain form of quotients:

$$\widetilde{\mathcal{U}} = \{(u^m)^* u^r \mid u \in A^*,\ m > 0,\ 0 \leqslant r < m\}$$

The study of these two classes is an intermediate step since:

$$\mathcal{B} \subseteq \widetilde{\mathcal{B}} \subseteq \mathcal{B}q \quad \text{and} \quad \mathcal{L} \subseteq \widetilde{\mathcal{L}} \subseteq \mathcal{L}q$$

Proposition 17. *Over a finite alphabet with at least two letters:*

1. *The class $\widetilde{\mathcal{B}}$ is defined by equations (E_1) and (E_6).*
2. *The class $\widetilde{\mathcal{L}}$ is defined by equations (E_1), (E_6), (E_5) and (E_7).*

From this proposition, we can see that the language presented in Example 4 $A^*(aa + bb)A^*$ is not in $\widetilde{\mathcal{B}}$, and therefore it is neither in $\mathcal{L}$ nor in $\mathcal{B}$, since the equation (E_6) is not satisfied. Indeed, it is sufficient to consider the word aba, and to observe that $(aba)^2aba \in A^*(aa + bb)A^*$ but $aba \notin A^*(aa + bb)A^*$.

As for the preceding cases, the languages in $\widetilde{\mathcal{B}}$ and $\widetilde{\mathcal{L}}$ can be written in a normal form: if L is a language in $\widetilde{\mathcal{B}}$ different from A^*, then

$$L \cup \{1\} = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i} \quad \text{or} \quad L - \{1\} = \left(\bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i} \right)^c$$

Similarly, if L is a language in $\widetilde{\mathcal{L}}$ different from A^*, then

$$L = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i}$$

where $(u_i)_{i=1,\dots,k}$ are words and m, $(r_i)_{i=1,\dots,k}$ are integers.

Proposition 18. *Let L be a regular language and m be the index of its syntactic monoid. The following properties are equivalent:*

1. *$L \in \widetilde{\mathcal{B}}$.*
2. *L satisfies the set of equations (E_1) and (E_6).*
3. *If $L \neq A^*$ then there are:*
   - *a finite set of words $(u_i)_{i=1\dots k}$,*
   - *a finite set of integers $(r_i)_{i=1\dots k}$ such that for all i, $0 \leqslant r_i < m$*

   such that:

$$L \cup \{1\} = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i} \quad \text{or} \quad L - \{1\} = (\bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i})^c$$

Proof. 1 *implies* 2*:* By Proposition 12, $(u^m)^*u^r$ satisfies (E_6). Moreover,

$$(u^m)^*u^r = (u^m)^*(u^{m-r})^{-1}$$

thus $(u^m)^*u^r \in \mathcal{B}q$ and thus by Theorem 2, it satisfies (E_1).

3 *implies* 1*:* The language $\{1\}$ belongs to $\widetilde{\mathcal{B}}$ since it is the intersection of $a^* \cap b^*$, a and b two different letters. Let L be a language such that

$$L \cup \{1\} = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i} \text{ or } L - \{1\} = (\bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i})^c$$

as given in point 3. By definition, the languages of the form $(u^m)^*u^r$ are in $\widetilde{\mathcal{B}}$. Finally, thanks to the closure by complement and by union we have $L \in \widetilde{\mathcal{B}}$.

2 *implies* 3*:* By Proposition 9.1 and 9.2, if a language L satisfies (E_1) and (E_6) then it also satisfies (E_2) and (E_3). Hence, by Theorem 2, L belongs to $\mathcal{B}q$. Therefore, if we denote by m the index of its syntactic monoid, there are:

- a finite set of words $(u_i)_{i=1\dots k}$,
- for all $i = 1, \dots, k$, a prefix of u_i denoted by p_i,
- a finite set of integers $(r_i)_{i=1\dots k}$ such that for all i, $r_i < 2m$,
- a finite set of words F,

Please cite this article in press as: L. Daviaud, C. Paperman, Classes of languages generated by the Kleene star of a word, Inf. Comput. (2018), https://doi.org/10.1016/j.ic.2018.07.002

such that:

$$L = (\bigcup_{i=1}^{k} u_i^{\geqslant r_i} p_i) \cup F \quad \text{or} \quad L = ((\bigcup_{i=1}^{k} u_i^{\geqslant r_i} p_i) \cup F)^c$$

Since $\widetilde{\mathcal{B}}$ is closed under complement, without loss of generality, we assume that:

$$L = (\bigcup_{i=1}^{k} u_i^{\geqslant r_i} p_i) \cup F$$

in the other case, we can argue on L^c. First, let $u \in F$. Then, $u \in \overline{L}$, and by (E_6), we have $u^{\omega+1} \in \overline{L}$. Therefore, $(u^m)^* u \in \overline{L}$. For this reason, we can assume that:

$$L = (\bigcup_{i=1}^{k} u_i^{\geqslant r_i} p_i) \cup \bigcup_{u \in F} (u^m)^* u \qquad (\star)$$

Moreover, assume we have a language of the form $u^{\geqslant r} p$ included in L. Then we have $u^\omega p \in \overline{L}$. Furthermore, by (E_6), $(u^\omega p)^{\omega+1} \in \overline{L}$ and we have:

$$(u^\omega p)^{\omega+1} = u^\omega (p u^\omega)^\omega p$$

Because $0 \notin \overline{L}$ and because $u^\omega (p u^m)^\omega \in \overline{L}$, we have $u^\omega (p u^m)^\omega \neq 0$. By the equation (E_1), there are a word t and two integers ℓ and k such that $u = t^k$ and $p u^m = t^\ell$. Therefore, we have $p = t^{\ell - km}$. Since p is a prefix of u, we necessarily have that $\ell - km \leqslant k$. Let $s = \ell - km$. Finally we have:

$$u^{\geqslant r} p = (t^k)^{\geqslant r} t^s \subseteq L$$

We define the following set:

$$R = \{r \in \mathbb{Z}/m\mathbb{Z} \mid \text{there is } n \text{ congruent to } r \text{ modulo } m \text{ such that } t^n \in (t^k)^{\geqslant r} t^s\}$$

Let $\ell \in R$ be different from 0. By definition, there exists an integer a such that $t^{am+\ell} \in L$. Since m is the idempotent power of L, we have $t^{\omega+\ell} \in \overline{L}$ and by equation (E_6), we have $(t^m)^* t^\ell \subseteq L$. If $0 \in R$ then we have $(t^m)^* \subseteq L \cup \{1\}$. Finally, we have that for each word u_i, there exist a word t_i, integers k_i, s_i and a set $R_i \subseteq \mathbb{Z}/m\mathbb{Z}$ such that

$$u_i^{\geqslant r_i} p_i = (t_i^{k_i})^{\geqslant r_i} t_i^{s_i} \subseteq \bigcup_{\ell \in R_i} (t_i^m)^* t_i^\ell \subseteq L \cup \{1\} \ .$$

By using equation $(\star)$, we obtain that:

$$\begin{aligned} L &= \bigcup_{u \in F} (u^m)^* u \cup \bigcup_{i=1}^{k} u_i^{\geqslant r_i} p_i \\ &\subseteq \bigcup_{u \in F} (u^m)^* u \cup \bigcup_{i=1}^{k} \bigcup_{\ell \in R_i} (t_i^m)^* t_i^\ell \\ &\subseteq L \cup \{1\} \end{aligned}$$

Finally, L has the form given in the proposition. □

Proposition 19. *Let L be a regular language and m be the index of its syntactic monoid. The following properties are equivalent:*

1. $L \in \widetilde{\mathcal{L}}$.
2. *L satisfies the set of equations (E_1), (E_6), (E_5) and (E_7).*
3. *If $L \neq A^*$ then there are:*
   - *a finite set of words $(u_i)_{i=1\ldots k}$,*
   - *a finite set of integers $(r_i)_{i=1\ldots k}$ such that for all i, $0 \leqslant r_i < m$*

   such that:

$$L = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i}$$

Proof. 1 *implies* 2*:* By Proposition 12, $(u^m)^*u^r$ satisfies (E_5) and (E_7). Moreover, $(u^m)^*u^r \in \widetilde{\mathcal{B}}$, thus by Proposition 18, it satisfies (E_1) and (E_6).

3 *implies* 1*:* By definition of $\widetilde{\mathcal{L}}$.

2 *implies* 3*:* Let L be a language different from A^* satisfying (E_1), (E_6) and (E_5). Let us denote by m the index of the syntactic monoid of L. First of all, by Proposition 18, since L satisfies (E_1) and (E_6) then there are:

- a finite set of words $(u_i)_{i=1\dots k}$,
- a finite set of integers $(r_i)_{i=1\dots k}$ such that for all i, $0 \leqslant r_i < m$,

such that:

$$L \cup \{1\} = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i} \quad \text{or} \quad L - \{1\} = \Big(\bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i}\Big)^c$$

The set of equations (E_5) means that if $0 \in \overline{L}$ then $L = A^*$. Indeed, for all words u, $0u = 0 \in \overline{L} \Rightarrow 1u = u \in \overline{L}$. So $0 \notin \overline{L}$ and thus we are in the case where:

$$L \cup \{1\} = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i}$$

If $1 \in L$ then $L = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i}$. Otherwise, there are two possibilities for 1 to be in $\bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i}$. First, if for some i, $r_i = 0$, then $(u_i^m)^{\geqslant 1} \subseteq L$ and by (E_7), 1 also belongs to L which is a contradiction. Secondly, if for some i, $u_i = 1$ we can just remove $(u_i^m)^* u_i^{r_i} = \{1\}$ from the second language to obtain the correct form:

$$L = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i} \quad \square$$

The case of $\mathcal{L}$

Finally, we can characterise the classes $\mathcal{L}$ and $\mathcal{B}$ by restricting the set of integers r_i that can be obtained in the normal form of $\widetilde{\mathcal{L}}$ and $\widetilde{\mathcal{B}}$. Regarding $\mathcal{L}$, one can prove that the only possible choice for r_i is 0. Thus, a language L different from A^* in $\mathcal{L}$ is of the form $L = \bigcup_{i=1}^{k} u_i^*$.

Unlike the class $\mathcal{L}$, the case of $\mathcal{B}$ can not be deduced directly from the case of $\widetilde{\mathcal{B}}$ and it is much more complicated. It is the subject of the next section.

Theorem 4. *Let L be a regular language. The following properties are equivalent:*

1. $L \in \mathcal{L}$.
2. *L satisfies the set of equations (E_1), (E_4) and (E_8).*
3. *If $L \neq A^*$ then there is a finite set of words $(u_i)_{i=1\dots k}$ such that:*

$$L = \bigcup_{i=1}^{k} u_i^*$$

Proof. 1 *implies* 2*:* By Proposition 10, u^* satisfies (E_1), (E_4), and (E_8).

3 *implies* 1*:* By definition of $\mathcal{L}q$.

2 *implies* 3*:* Let L be a language different from A^* satisfying (E_1), (E_4) and (E_8). Let us denote by m the index of its syntactic monoid and by φ its syntactic morphism. By Propositions 9.3, 9.4 and 9.5, L also satisfies equations (E_5), (E_6) and (E_7). By Proposition 19, since L satisfies (E_1), (E_5), (E_6) and (E_7) then there are:

- a finite set of words $(u_i)_{i=1\dots k}$,
- a finite set of integers $(r_i)_{i=1\dots k}$ such that for all i, $0 \leqslant r_i < m$,

such that:

$$L = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i}$$

Please cite this article in press as: L. Daviaud, C. Paperman, Classes of languages generated by the Kleene star of a word, Inf. Comput. (2018), https://doi.org/10.1016/j.ic.2018.07.002

Consider $i \in \{1, \dots, k\}$ and let $d_i = \gcd(m, r_i)$. Since $u_i^{r_i}$ belongs to L we have, by equation (E_8), that $(u_i^{r_i})^* \subseteq L$. There exists a positive integer b such that br_i is congruent to d_i modulo m. Because m is the index of the syntactic monoid of L, we have:

$$\varphi(u_i^{br_i}) = \varphi(u_i^{m+d_i}) = \widehat{\varphi}(u_i^{\omega+d_i})$$

Thus, $u_i^{\omega} u_i^{d_i} \in \overline{L}$ and, by equation (E_4), $u_i^{d_i} \in L$. By applying equation (E_8), we have that $(u_i^{d_i})^* \subseteq L$, and finally we can conclude since:

$$L = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i} \subseteq \bigcup_{i=1}^{k} (u_i^{d_i})^* \subseteq L \quad \square$$

5. The case of the Boolean algebra $\mathcal{B}$

We enter here the most intricate part of the description of the classes generated by the languages of the form u^*. The idea is to restrict the possible integers r_i we can obtain in the description of $\widetilde{\mathcal{B}}$. For that, we will define equivalence relations over the integers. Once this will be done, the main difficulty will be to translate properties over integers into profinite equations. In order to do that, we will introduce profinite numbers. This issue is addressed in Section 5.1 which first defines which sets of integers are allowed for the r_i and then translates it into equations. Finally, Section 5.2 aggregates all these notions to give the characterisation of $\mathcal{B}$.

5.1. Equivalence classes over $\mathbb{N}$ and profinite numbers

Let m be an integer, and r and s be in $\{0, \dots, m-1\}$, let us define $r \equiv_m s$ if and only if $\gcd(r, m) = \gcd(s, m)$. Observe that $\equiv_m$ is an equivalence relation. Intuitively, a language in $\mathcal{B}$ with m as the index of its syntactic monoid, will not be able to separate two integers that are equivalent with respect to $\equiv_m$. More precisely, let L be a language in $\mathcal{B}$ with m as the index of its syntactic monoid and $r \equiv_m s$. Then for all words u and for all k, k', we have $u^{km+r} \in L$ if and only if $u^{k'm+s} \in L$.

Example 5. We introduce the language $L = (a^2)^* - (a^6)^*$. This language is, by definition, in $\mathcal{B}$. The index of its syntactic monoid is 6. Classes for $\equiv_6$ are $\{1, 5\}$, $\{2, 4\}$, $\{3\}$ and $\{0\}$. Thus, L cannot separate a word in $(a^6)^* a^2$ from a word in $(a^6)^* a^4$. Therefore, since $(a^6)^* a^2$ is in L, $(a^6)^* a^4$ is also in L. Since L belongs to $\mathcal{B}$, it also belongs to $\widetilde{\mathcal{B}}$ and we have a convenient normal form given by Proposition 17:

$$L = (a^2)^* - (a^6)^* = (a^6)^* a^2 \cup (a^6)^* a^4$$

The equivalence relation $\equiv_m$ allows us to give the form of the languages in $\mathcal{B}$. The next step is to translate it in terms of equations. The difficulty comes from the fact that $\equiv_m$ depends on the parameter m that represents the index of the syntactic monoid of a given language. So, this cannot be directly translated into a set of equations that are supposed to not depend on a specific language.

Profinite numbers. Consider a one-letter alphabet $B = \{a\}$ and the profinite monoid $\widehat{B^*}$. There is an isomorphism from B^* to $\mathbb{N}$ that associates a word to its length. Then there is a unique set $\widehat{\mathbb{N}}$ and a unique isomorphism $\psi : \widehat{B^*} \to \widehat{\mathbb{N}}$ such that $\mathbb{N} \subseteq \widehat{\mathbb{N}}$ and $\widehat{\psi}$ coincides with ψ on $\mathbb{N}$. Elements of $\widehat{\mathbb{N}}$ are called *profinite numbers*. They are limits of sequences of integers, in the sense of the topology of the set of words on a one-letter alphabet. Given a word u, and a profinite number α, u^α corresponds to the profinite word that is the limit of the words u^{α_n} where $(\alpha_n)_n$ is a sequence of integers converging to α.

Let $\mathcal{P} = \{p_1 < p_2 < \ldots < p_n < \ldots\}$ be a cofinite sequence of prime numbers. That is, a sequence of prime numbers such that only a finite number of prime numbers are not used in the sequence. Consider the sequence defined by $z_n^{\mathcal{P}} = (p_1 \cdots p_n)^{n!}$. The sequence $(z_n^{\mathcal{P}})_{n>0}$ is converging in $\widehat{\mathbb{N}}$ and we denote by $z^{\mathcal{P}}$ its limit.

We can give now the last set of equations needed to characterise $\mathcal{B}$ and that conveys the notion of equivalence over $\mathbb{N}$ defined above. Denote by Γ the set of pairs of profinite numbers $(dz^{\mathcal{P}}, dpz^{\mathcal{P}})$ satisfying the three following conditions:

- $\mathcal{P}$ is a cofinite sequence of prime numbers,
- $p \in \mathcal{P}$,
- if q divides d then $q \notin \mathcal{P}$.

Let us define the set of equations (E_9) by:

$$x^\alpha \leftrightarrow x^\beta \text{ for all } (\alpha, \beta) \in \Gamma \qquad (E_9)$$

We first prove that a language u^* satisfies this set of equations. Let us denote by $\overline{d}^m$ the equivalence class of d for $\equiv_m$.

Proposition 20. *Let u be a primitive word and m be an integer, then $(u^m)^*$ satisfies the set of equations (E_9).*

Proof. Let $(dz^{\mathcal{P}}, dpz^{\mathcal{P}})$ be in Γ and x be a word. Let $L = (u^m)^*$. We suppose that $x^{dz^{\mathcal{P}}} \in \overline{L}$. Then, there is k such that $x = u^k$ and moreover for all n large enough, $kdz_n^{\mathcal{P}}$ is a multiple of m. Thus $kdpz_n^{\mathcal{P}}$ is also a multiple of m, and thus $x^{dpz^{\mathcal{P}}} \in \overline{L}$. Conversely, suppose that $x^{dpz^{\mathcal{P}}} \in \overline{L}$, then there is k such that $x = u^k$ and moreover for all n large enough, $kdpz_n^{\mathcal{P}}$ is a multiple of m. Finally, $kdz_{n+1}^{\mathcal{P}}$ is a multiple of $kdz_n^{\mathcal{P}} p^{n+1}$ since $p \in \mathcal{P}$, that is a multiple of $kdpz_n^{\mathcal{P}}$ that is a multiple of m. Thus, for n large enough, $kdz_{n+1}^{\mathcal{P}}$ is a multiple of m, and $x^{dpz^{\mathcal{P}}} \in \overline{L}$. □

5.2. Characterisation of $\mathcal{B}$

The following result combines the notions given in Section 5.1 and characterises the class $\mathcal{B}$.

Theorem 5. *Over a finite alphabet with at least two letters, the class $\mathcal{B}$ is defined by equations (E_1), (E_6) and (E_9).*

Let us first give a sketch of the proof. First, we have proven that u^* satisfies (E_1), (E_6) and (E_9).

The reverse implication is proved in two steps. First, we prove that if a language L is different from A^* and satisfies (E_1), (E_6) and (E_9), then just like for the other classes, it has a normal form:

$$L \cup \{1\} = \bigcup_{i=1}^{k} \bigcup_{r \in S_i} (u_i^m)^* u_i^r \quad \text{or} \quad \big(L - \{1\}\big)^c = \bigcup_{i=1}^{k} \bigcup_{r \in S_i} (u_i^m)^* u_i^r$$

where m is an integer, $(u_i)_{i=1,\dots,k}$ is a finite set of words, and S_i is an equivalence class of $\equiv_m$. We start by using the first part of Proposition 17 to prove that L belongs to $\widetilde{\mathcal{B}}$. So L can be written as:

$$L \cup \{1\} = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i} \quad \text{or} \quad L - \{1\} = \big(\bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i}\big)^c$$

We prove that for all $t \equiv_m r$, u^r belongs to L if and only if u^t belongs to L. The idea is the following: Let φ be the syntactic morphism of L, consider any cofinite sequence of prime numbers $\mathcal{P}$. If all the prime divisors of m are in $\mathcal{P}$, then for all n large enough, m divides $z_n^{\mathcal{P}}$ and thus for all words x, $\widehat{\varphi}(x^{dz^{\mathcal{P}}}) = \widehat{\varphi}(x^{\omega}) = \widehat{\varphi}(x^{dpz^{\mathcal{P}}})$. If none of the prime divisors of m is in $\mathcal{P}$, then for all n large enough, $z_n^{\mathcal{P}}$ is of the form $km + 1$. Then $\widehat{\varphi}(x^{dz^{\mathcal{P}}}) = \widehat{\varphi}(x^{\omega+d})$ and $\widehat{\varphi}(x^{dpz^{\mathcal{P}}}) = \widehat{\varphi}(x^{\omega+dp})$. Finally, d and dp under the conditions that define the set Γ, represent integers in the same equivalence class with respect to m that are then linked by (E_9). Other situations are combinations of these two.

Once we have the normal form for L, what is left is to prove that a language that can be written in this normal form belongs to $\mathcal{B}$. This is done by proving that:

$$\bigcup_{p \in \overline{r}^m} (u^m)^* u^p = (u^d)^* - \bigcup_{\substack{k \text{ s.t.} \\ 0 \leqslant k \leqslant m \\ gcd(k, \frac{m}{d}) \neq 1}} (u^{kd})^*$$

where $\overline{r}^m$ is the equivalence class of r for $\equiv_m$ and $d = \gcd(m, r)$.

Now let us give the complete proof.

Proposition 21. *Let L be a language and m be the index of its syntactic monoid. Suppose that there are an integer k and pairwise distinct words $u_1, \dots, u_k$ such that:*

$$L = \bigcup_{i=1}^{k} \bigcup_{r \in S_i} (u_i^m)^* u_i^r$$

where for all i, $S_i \subseteq \{0, \dots m-1\}$. Moreover suppose that L satisfies (E_6) and (E_9). Then, for all i and all $r \in S_i$, one has $\overline{r}^m \subseteq S_i$.

Proof. Let $i \in \{1, \dots, k\}$ and $r \in S_i$, let us prove that $\overline{r}^m \subseteq S_i$. First, if $r = 0$ then $\overline{r}^m = \{0\} \subseteq S_i$. Otherwise, let $d = \gcd(r, m)$. By definition,

$$\overline{r}^m = \{d\ell \mid 0 < \ell < \frac{m}{d},\ \gcd(\ell, \frac{m}{d}) = 1\}$$

Let $0 < \ell < \frac{m}{d}$ be such that $\gcd(\ell, \frac{m}{d}) = 1$. We will use the following theorem:

Theorem 6 *(Dirichlet's theorem on arithmetic progressions). For all positive integers a, b such that* $\gcd(a,b)=1$*, there are infinitely many prime numbers congruent to a modulo b.*

By this theorem, there is a prime number $p > m$ such that p is congruent to ℓ modulo $\frac{m}{d}$. Consequently, dp is congruent to $d\ell$ modulo m. Let us consider $\mathcal{P}$ the cofinite sequence of prime numbers that do not divide m. In particular, $p \in \mathcal{P}$ and for all q that divide d, q does not belong to $\mathcal{P}$. Moreover, for n large enough, $z_n^{\mathcal{P}}$ is congruent to 1 modulo m. Indeed, by definition, a prime number $p \in \mathcal{P}$ does not divide m and thus $\gcd(m,p)=1$. Then there exist $n' < m$ such that $p^{n'}$ is congruent to 1 modulo m (n' is the order of the cyclic group formed by the elements t modulo m, such that $\gcd(t,m)=1$). Consider $n > m!$ then for all $p \in \mathcal{P}$, p^n is congruent to 1 modulo m and so is $z_n^{\mathcal{P}}$. Then one gets:

$$\begin{aligned}
u_i^d &\leftrightarrow u_i^{\omega+d} && \text{(by equation } (E_6)\text{)}\\
&\leftrightarrow u_i^{dz^{\mathcal{P}}}\\
&\leftrightarrow u_i^{dpz^{\mathcal{P}}} && \text{(by equation } (E_9)\text{)}\\
&\leftrightarrow u_i^{\omega+d\ell}\\
&\leftrightarrow u_i^{d\ell} && \text{(by equation } (E_6)\text{)}
\end{aligned}$$

Hence, since $r \in S_i$ then $d \in S_i$ and finally $\overline{r}^m \subseteq S_i$. □

Proposition 22. *Let u be a word and* $r < m$ *be integers. Then:*

$$\bigcup_{p\in\overline{r}^m} (u^m)^* u^p \in \mathcal{B}$$

Proof. By definition, there is d such that for all $p \in \overline{r}^m$, $\gcd(p,m)=d$. Let us show that:

$$\bigcup_{p\in\overline{r}^m} (u^m)^* u^p = (u^d)^* - \bigcup_{\substack{k \text{ s.t.}\\ 0\leqslant k\leqslant m\\ gcd(k,\frac{m}{d})\neq 1}} (u^{kd})^*$$

which will conclude the proof. For the first inclusion, let $p \in \overline{r}^m$, and let ℓ be an integer. We have the following:

- the integer d divides $\ell m + p$. Thus $u^{\ell m+p} \in (u^d)^*$,
- let k be an integer such that $\gcd(k,\frac{m}{d}) = c \neq 1$. The integer c divides both k and $\frac{m}{d}$. If k divides $\ell\frac{m}{d} + \frac{p}{d}$ then necessarily c divides also $\frac{p}{d}$. But $\gcd(\frac{m}{d}, \frac{p}{d}) = 1$. Hence this is impossible. So k does not divide $\ell\frac{m}{d} + \frac{p}{d}$, and so kd does not divide $\ell m + p$ and finally

$$u^{\ell m+p} \notin \bigcup_{\substack{k \text{ s.t.}\\ 0\leqslant k\leqslant m\\ gcd(k,\frac{m}{d})\neq 1}} (u^{kd})^*$$

That is why:

$$u^{\ell m+p} \in (u^d)^* - \bigcup_{\substack{k \text{ s.t.}\\ 0\leqslant k\leqslant m\\ \gcd(k,\frac{m}{d})\neq 1}} (u^{kd})^*$$

As for the reverse inclusion, let ℓ be an integer such that:

$$u^{d\ell} \in (u^d)^* - \bigcup_{\substack{k \text{ s.t.}\\ 0\leqslant k\leqslant m\\ \gcd(k,\frac{m}{d})\neq 1}} (u^{kd})^* \qquad (\star)$$

One has $\ell d = qm + p$ for some $0 \leqslant p < m$ (euclidean division). It remains to prove that this p belongs to $\overline{r}^m$ that is to say that $\gcd(p,m)=d$. First d divides m and then p. Then one gets $\ell = q\frac{m}{d} + \frac{p}{d}$. So, a divisor of $\frac{p}{d}$ and $\frac{m}{d}$ divides ℓ. Let k be such an integer then by $(\star)$, either $k > m$ or $\gcd(k,\frac{m}{d}) = 1$ and then $k = 1$. So the only divisor of $\frac{p}{d}$ and $\frac{m}{d}$ is 1, and so we get $\gcd(p,m)=d$ and then

Please cite this article in press as: L. Daviaud, C. Paperman, Classes of languages generated by the Kleene star of a word, Inf. Comput. (2018), https://doi.org/10.1016/j.ic.2018.07.002

$$u^{d\ell} \in \bigcup_{p\in\overline{r}^m} (u^m)^* u^p \quad \square$$

Theorem 7. *Let L be a regular language and m be the index of its syntactic monoid. The following properties are equivalent:*

1. $L \in \mathcal{B}$.
2. *L satisfies the sets of equations (E_1), (E_6) and (E_9).*
3. *If $L \neq A^*$ then there are:*
   - *a finite set of pairwise distinct words $(u_i)_{i=1\dots k}$,*
   - *for each i, a finite set of integers S_i that is a finite union of equivalence classes of $\equiv_m$,*

   such that:

$$L \cup \{1\} = \bigcup_{i=1}^{k} \bigcup_{r\in S_i} (u_i^m)^* u_i^r \quad \text{or} \quad L - \{1\} = (\bigcup_{i=1}^{k} \bigcup_{r\in S_i} (u_i^m)^* u_i^r)^c$$

Proof. 1 *implies* 2*:* By Proposition 10, u^* satisfies the sets of equations (E_1), (E_6) and (E_9).

2 *implies* 3*:* Suppose that L satisfies (E_1), (E_6) and (E_9) and is different from A^*. By Theorem 2, there is an integer $k \geqslant 1$ and for all $i \in \{1, \dots, k\}$, a word u_i and a non-negative integer $r_i \leqslant m$ such that:

$$L = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i} \quad \text{or} \quad L = (\bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i})^c$$

where m is the index of the syntactic monoid of L. Suppose that we are in the first case and that $L = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i}$, otherwise we can argue on L^c that satisfies the same equations. One can reformulate: there are an integer $k \geqslant 1$, a finite sequence of pairwise distinct words $(u_i)_{i=1,\dots,k}$ and sets of integers $S_i \subseteq \{1, \dots, m\}$ such that: $L = \bigcup_{i=1}^{k} \bigcup_{r\in S_i} (u_i^m)^* u_i^r$. Since L satisfies (E_6) and (E_9) then by Proposition 21, for all i and all $r \in S_i$, $\overline{r}^m \subseteq S_i$. Thus each S_i is a finite union of equivalence classes which concludes the proof.

3 *implies* 1*:* By Proposition 22 and the fact that $\mathcal{B}$ is closed under union. $\square$

6. Decidability

The characterisations that are given in Theorems 1 and 5 yield as a counterpart the decidability of the classes $\mathcal{B}q$, $\mathcal{L}q$, $\mathcal{L}$ and $\mathcal{B}$: given a regular language L, one can decide if L belongs to said classes. Every single equation is effectively testable. The main issue is to test an infinite set of equations in finite time. The idea is to test the equations in the syntactic monoid of L that is finite and thus test a finite number of equations. The first step is to compute M, the syntactic monoid of L, m its index, φ the syntactic morphism and P, the syntactic image of L. They are all computable from the minimal automaton of L. Then, it is sufficient to check if the sets of equations are satisfied directly in M and P, which are finite. More precisely:
(E_4): for all $x, y, z \in M$, $yx^m z \in P \Rightarrow yz \in P$
(E_5): particular case of (E_4)
(E_6): for all $x \in M$, for all $0 < \ell < m$, $x^\ell \in P \Leftrightarrow x^{m+\ell} \in P$
(E_7): particular case of (E_4)
(E_8): for all $x \in M$, for all $0 < \ell \leqslant 2m$, $x \in P \Rightarrow x^\ell \in P$
(E_9): thanks to the notion of equivalence classes given in Section 5.1, testing equations in (E_9) is the same as testing that for all $x \in M$, for all $0 \leqslant r, s < m$ such that $r \equiv_m s$, $x^r \in P \Leftrightarrow x^s \in P$.

It is much more difficult to translate sets of equations (E_1), (E_2) and (E_3) in M since conditions "$xy \neq yx$", "$y \notin P_x$" and "$y \notin S_x$" cannot be translated directly in M.
(E_1): consider $x, y \in M$ such that $x^m y^m \neq 0$. One has to check that for all words $u \in \varphi^{-1}(x)$, $v \in \varphi^{-1}(y)$, $uv = vu$.
(E_2): consider $x, y \in M$ such that $x^m y \neq 0$. One has to check that for all words $u \in \varphi^{-1}(x)$, $v \in \varphi^{-1}(y)$, $v \in P_u$.
(E_3): same as (E_2)

Let us now prove the decidability of the sets of equations (E_1) and (E_2). We are going to prove that one can decide if a given language L satisfies these sets of equations. Consider a non empty regular language L, its syntactic monoid M, its syntactic morphism φ and its syntactic image P. We will use the following basic lemma, which can be found for example in [7]:

Lemma 3. *For all words $u, v \in A^*$, if $uv = vu$ then there is a primitive word $w \in A^*$ such that $u \in t^*$ and $v \in t^*$.*

First let us remind (E_1):

$$x^\omega y^\omega = 0 \text{ for } x, y \in A^* \text{ such that } xy \neq yx \qquad (E_1)$$

Please cite this article in press as: L. Daviaud, C. Paperman, Classes of languages generated by the Kleene star of a word, Inf. Comput. (2018), https://doi.org/10.1016/j.ic.2018.07.002

The language L satisfies (E_1) if and only if for all $x, y \in M$ such that $x^\omega y^\omega \neq 0$, for all words $u \in \varphi^{-1}(x)$, $v \in \varphi^{-1}(y)$, one has $uv = vu$. For proving the decidability of this problem, it is sufficient to prove the decidability of the following one:

Given two non empty regular languages L and K, for all words $u \in L$ and $v \in K$, are uv and vu equal?

Moreover, as shown by the following lemma, this problem is equivalent to checking if for all words $u \in L \cup K$ and $v \in L \cup K$, $uv = vu$.

Lemma 4. *For all words u, v, w with $w \neq 1$, if $uw = wu$ and $vw = wv$ then $uv = vu$.*

Proof. Suppose that $uw = wu$ and $vw = wv$. Then, by using Lemma 3, there are two primitive words $t, s \in A^*$ such that $u, w \in t^*$ and $v, w \in s^*$. Thus $w \in t^* \cap s^*$. But since s and t are primitive, necessarily $s = t$. Finally, $u, v, w \in t^* = s^*$ and $uv = vu$. □

Thus, it is sufficient to consider the problem:

Given a regular language L, for all words $u, v \in L$, $uv = vu$?

Finally, we prove that a language satisfies this property if and only if it is a subset of a language t^* for a word t. This property is decidable: consider the smallest word $s \in L$ with respect to the shortlex order and t the (unique) primitive word such that $s \in t^*$. It is sufficient to check that $L \subseteq t^*$.

Lemma 5. *Let L be a regular language. For all $u, v \in L$, $uv = vu$ if and only if there is $t \in A^*$ such that $L \subseteq t^*$.*

Proof. $\Leftarrow$: clear

$\Rightarrow$: Let $u \in L - \{1\}$, denote by t_u the (unique) primitive word such that $u \in t_u^*$. Let $v \in L$, since $uv = vu$ then, thanks to Lemma 3, there is a primitive word t such that $u, v \in t^*$. Necessarily, $t = t_u = t_v$. Thus, $L \subseteq t^*$. □

Let us now remind the equation (E_2):

$$x^\omega y = 0 \text{ for } x, y \in A^* \text{ such that } y \notin P_x \qquad (E_2)$$

The language L satisfies (E_2) if and only if for all $x, y \in M$ such that $x^\omega y \neq 0$, for all words $u \in \varphi^{-1}(x)$, $v \in \varphi^{-1}(y)$, one has $v \in P_u$. For proving the decidability of this problem, it is sufficient to prove the decidability of the following one:

Given two regular languages L and K, for all words $u \in L$ and $v \in K$, does v belong to P_u?

We prove that this problem is decidable by distinguishing two cases. First, we deal with the case where K is finite. Let $v \in K$. For all $u \in L$ such that $|u| \leqslant |v|$, check that $v \in P_u$. Otherwise, if $|u| > |v|$, the condition $v \in P_u$ means that v is a prefix of u. So, one can just test if $\{u \in L \mid |u| > |v|\}$ is a subset of vA^*. Second, we assume that K is infinite. Then, as shown in the next lemma, L satisfies that for every $u, w \in L$ we have $uw = wu$.

Lemma 6. *Let u, v, w words such that $v \in P_u$, $v \in P_w$ and $|v| > 2\max(|u|, |w|)$. Then, $uw = wu$.*

Proof. Let $v \in P_u \cap P_w$. Then there are two non negative integers ℓ, k, a prefix of u, p_u and a prefix of w, p_w such that $v = u^\ell p_u = w^k p_w$ $(\star)$. Since $|v| \geqslant 2\max(|u|, |w|)$ then $\min(\ell, k) \geqslant 2$. Let us suppose that $\ell \leqslant k$ (and thus $|w| \leqslant |u|$). By $(\star)$ and since $\ell \geqslant 2$, there are two words w_1, w_2 and a non negative integer p such that:

- $w = w_1 w_2$,
- $u = (w_1 w_2)^p w_1$,
- $w_2 w_1$ is a prefix of u.

Then, $w_1 w_2 = w_2 w_1$, and $uw = wu$. □

Moreover, thanks to Lemma 5, a language L satisfies for every $u, w \in L$, $uw = wu$ if and only if it is a subset of a language t^* for a word t. The first step is to check that $L \subseteq t^*$ and to compute such a word t. Finally, if $L \subseteq t^*$, the two following conditions are clearly equivalent:

1. For all words $u \in L$ and $v \in K$, $v \in P_u$.
2. For all words $v \in K$, $v \in P_t$.

Please cite this article in press as: L. Daviaud, C. Paperman, Classes of languages generated by the Kleene star of a word, Inf. Comput. (2018), https://doi.org/10.1016/j.ic.2018.07.002

Thus, it is now sufficient to check that $K \subseteq P_t$, which is decidable.

7. The case of a unary alphabet

This section summarises results for a unary alphabet. In this case, the syntactic monoid of a language of the form $(a^k)^*$ has no zero and moreover the construction of ρ_A, given in [1,12] for larger alphabets, does not make any sense for a singleton alphabet. To extend the proof of the two-letter case, we will use the fact that a regular language over the alphabet $A = \{a\}$ is a finite union of languages of the form $(a^q)^*a^p$ for non negative integers p and q. Observe that finite languages are those with $q = 0$. We can derive from proofs made for the general case the form of languages and the equations characterising $\mathcal{B}q$, $\mathcal{L}q$, $\mathcal{L}$ and $\mathcal{B}$. The one letter case $\mathcal{B}q$, $\mathcal{L}q$, $\mathcal{L}$ and $\mathcal{B}$ could be adapted from the proof but is not, unfortunately, a straightforward consequence. In particular, it is needed to redo some of the proof.

The major difference with the two-letter case is that $\mathcal{B}q$ can no longer capture the finite languages (except the empty language and the language containing only the empty word). We will show in Theorem 8 that it is characterised by the equations:

$$x^{\omega+1} = x \tag{E_{10}}$$

Moreover it contains languages that are a finite union of languages of the form $(a^q)^*a^p$ for $q > 0$ and $0 \leqslant p \leqslant q$ (plus the language $\{1\}$).

Classes $\mathcal{L}q$, $\mathcal{L}$ and $\mathcal{B}$ are restrictions of $\mathcal{B}q$ in the same way as in the two-letter case. In Theorem 8, we will prove the following characterisations.

- Languages in $\mathcal{L}q$ are a finite union of languages of the form $(a^q)^*a^p$ for $q > 0$ and $0 \leqslant p < q$ (plus the language $\{1\}$). The class $\mathcal{L}q$ is characterised by equations (E_4) and (E_{10}).
- Languages in $\mathcal{L}$ are a finite union of languages of the form $(a^q)^*$ for a non negative integer q, and $\mathcal{L}$ is characterised by equations (E_4), (E_8) and (E_{10}).
- Languages in $\mathcal{B}$ are a finite union of languages of the form $\cup_{p\in S}(a^q)^*a^p$ for a positive integer q and S an equivalence class of $\equiv_q$ and languages of the form $(a^q)^{\geqslant 1}$ for a non negative integer q. The class $\mathcal{B}$ is characterised by equations (E_9) and (E_{10}).

The syntactic monoid of $(a^q)^*$, for a positive integer q, is the cyclic group $\mathbb{Z}/q\mathbb{Z}$. We will use the multiplicative notations: $1, a, \dots a^{q-1}$. The syntactic morphism associates a word a^p to $a^{p \bmod q}$. Thus the only idempotent in the syntactic monoid of $(a^q)^*$ is the image of the empty word 1 whose preimage is the language $(a^q)^*$.

The syntactic monoid of $\{1\}$ contains two elements, 1 and 0, both idempotent. The element 1 is the image of the empty word, while 0 is a zero and is the image of all non empty words.

Proposition 23. *For all non negative integers q, the language $(a^q)^*$ satisfies the sets of equations (E_4), (E_8), (E_9), (E_{10}).*

Proof. (E_4): It is the same proof as in the two-letter case given in Proposition 10.

(E_8): By definition. It is the same proof as in the two-letter case given in Proposition 10.

(E_9): It is the same proof as in the two-letter case given in Proposition 20.

(E_{10}): If $q = 0$, that is to say if we consider the language $\{1\}$, the two elements of the syntactic monoid satisfy the equation since $0^{\omega+1} = 0$ and $1^{\omega+1} = 1$. If $q \neq 0$, then the syntactic monoid of $(a^q)^*$ contains only one idempotent (the image of 1, or a^q). If x is an element of the cyclic group, then $x^{\omega+1} = 1x = x$. □

Proposition 24. *For all $q > 0$, $0 \leqslant p \leqslant q$, the language $(a^q)^*a^p$ belongs to $\mathcal{B}q$.*

Proof. If $0 \leqslant p < q$, then $(a^q)^*a^p = (a^q)^*(a^{q-p})^{-1}$, thus belongs to $\mathcal{B}q$. If $p = q$, then $(a^q)^*a^p = (a^q)^+ = (a^q)^* - \{1\} \in \mathcal{B}q$. □

Theorem 8. *Over a unary alphabet:*

1. *The class $\mathcal{B}q$ is defined by equations (E_{10}).*
2. *The class $\mathcal{L}q$ is defined by equations (E_4) and (E_{10}).*
3. *The class $\mathcal{L}$ is defined by equations (E_4), (E_8) and (E_{10}).*
4. *The class $\mathcal{B}$ is defined by equations (E_9) and (E_{10}).*

Proof. By Proposition 23, the languages that generate respectively $\mathcal{B}q$, $\mathcal{L}q$, $\mathcal{L}$ and $\mathcal{B}$ satisfy (E_4), (E_8), (E_9), (E_{10}). What remains to be done is to prove that a regular language that satisfy (E_{10}) (resp. plus (E_4), (E_8), (E_9)) belongs to $\mathcal{B}q$ (resp. $\mathcal{L}q$, $\mathcal{L}$, $\mathcal{B}$).

1. We will use the well-known form of regular languages over a one-letter alphabet, given in the following Lemma.

Please cite this article in press as: L. Daviaud, C. Paperman, Classes of languages generated by the Kleene star of a word, Inf. Comput. (2018), https://doi.org/10.1016/j.ic.2018.07.002

Lemma 7 *(Folklore). A regular language over the alphabet $\{a\}$ is a finite union of languages of the form $(a^q)^*a^p$ for two non negative integers q and p.*

Let L be a regular language satisfying (E_{10}). Let m denote the index of the syntactic monoid of L. First, for $p > 0$, if $a^p \in L$ then $(a^m)^*a^p \in L$ thanks to (E_{10}). Then by using Lemma 7, L is a finite union of languages of the form $(a^q)^*a^p$ for $q > 0$ and $p \geqslant 0$ and of the language $\{1\}$. Now, suppose that $(a^q)^*a^p \subseteq L$ for some $q > 0$ and $p > q$, then $a^{2mq}a^{p-q} \in L$ and by (E_{10}), $a^{p-q} \in L$. Thus, $(a^q)^*a^{p-q} \subseteq L$. Thus, finally L is a finite union of languages of the form $(a^q)^*a^p$ for $q > 0$ and $0 \leqslant p \leqslant q$ and of the language $\{1\}$. We can now conclude by using Proposition 24 that $L \in \mathcal{B}q$.

2. Let L be a regular language satisfying (E_{10}) and (E_4). By the previous item, $L \in \mathcal{B}q$, thus is a finite union of languages of the form $(a^q)^*a^p$ for $q > 0$ and $0 \leqslant p \leqslant q$ and of the language $\{1\}$. Let $q > 0$. Suppose that $(a^q)^{\geqslant 1} \subseteq L$. Then by ($E_4$), $1 \in L$, and thus $(a^q)^* \subseteq L$. So, finally, L is a finite union of languages of the form $(a^q)^*a^p$ for $q > 0$ and $0 \leqslant p < q$ and of the language $\{1\}$. And for all $q > 0$ and $0 \leqslant p < q$, $(a^q)^*a^p = (a^q)^*(a^{q-p})^{-1} \in \mathcal{L}q$. Thus $L \in \mathcal{L}q$.

3. and 4. The proofs of these two items are similar to their generic case. The proof of item 3, is the same as the one given in Theorem 4. As for the proof of item 4., it is similar to the one given in Theorem 7 in the two-letter case. □

8. Conclusion

This paper gives an equational description of the lattice and Boolean algebra generated by languages of the form u^*; as well as of their closures under quotients. These descriptions illustrate the power of the topological framework introduced by [5]. In particular, it gives us tools to describe in an effective way these classes of languages.

The next step could be to investigate either the case of the classes of languages generated by F^* where F is a finite set of words, or the case of the classes generated by $u_1^*u_2^*\dots u_k^*$ where $u_1, \dots, u_k$ are words. An answer to either of these questions will help us gain a better understanding of the phenomena that appear in the study of the variety generated by the languages u^* and of the generalised star-height problem.

Acknowledgments

We would like to thank Olivier Carton for numerous enlightening discussions and Jean-Éric Pin for suggesting this interesting question to us and supporting us in our way to the solution.

References

[1] J. Almeida, M.V. Volkov, Profinite identities for finite semigroups whose subgroups belong to a given pseudovariety, J. Algebra Appl. 2 (2) (2003) 137–163.
[2] J. Almeida, P. Weil, Relatively free profinite monoids: an introduction and examples, in: Semigroups, Formal Languages and Groups, York, 1993, in: NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., vol. 466, Kluwer Acad. Publ., Dordrecht, 1995, pp. 73–117.
[3] L. Daviaud, C. Paperman, Classes of languages generated by the Kleene star of a word, in: Mathematical Foundations of Computer Science 2015–40th International Symposium, Proceedings, Part I, MFCS 2015, Milan, Italy, August 24–28, 2015, 2015, pp. 167–178.
[4] S. Eilenberg, Schützenberger, On pseudovarieties, Adv. Math. 19 (1976) 413–418.
[5] M. Gehrke, S. Grigorieff, J-É. Pin, Duality and equational theory of regular languages, in: L. Aceto, et al. (Eds.), ICALP 2008, Part II, in: Lect. Notes Comp. Sci., vol. 5126, Springer, Berlin, 2008, pp. 246–257.
[6] J. Honkala, On slender languages, in: Current Trends in Theoretical Computer Science, World Sci. Publ., River Edge, NJ, 2001, pp. 708–716.
[7] M. Lothaire, Combinatorics on Words, Cambridge Mathematical Library, Cambridge University Press, Cambridge, 1997, With a foreword by Roger Lyndon and a preface by Dominique Perrin, Corrected reprint of the 1983 original, with a new preface by Perrin.
[8] R. McNaughton, S. Papert, Counter-Free Automata, The M.I.T. Press, Cambridge, Mass.-London, 1971.
[9] J-É. Pin, Mathematical foundations of automata theory.
[10] J-É. Pin, Profinite methods in automata theory, in: Susanne Albers, Jean-Yves Marion (Eds.), 26th International Symposium on Theoretical Aspects of Computer Science, STACS 2009, Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, 2009, pp. 31–50.
[11] J-É. Pin, H. Straubing, D. Thérien, Some results on the generalized star-height problem, Inf. Comput. 101 (1992) 219–250.
[12] N.R. Reilly, S. Zhang, Decomposition of the lattice of pseudovarieties of finite semigroups induced by bands, Algebra Univers. 44 (3–4) (2000) 217–239.
[13] J. Reiterman, The Birkhoff theorem for finite algebras, Algebra Univers. 14 (1) (1982) 1–10.
[14] M.P. Schützenberger, On finite monoids having only trivial subgroups, Inf. Control 8 (1965) 190–194.
[15] S. Yu, Regular languages, in: G. Rozenberg, A. Salomaa (Eds.), Handbook of Language Theory, vol. 1, Springer, 1997, pp. 679–746, chapter 2.

Please cite this article in press as: L. Daviaud, C. Paperman, Classes of languages generated by the Kleene star of a word, Inf. Comput. (2018), https://doi.org/10.1016/j.ic.2018.07.002