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

A R T I C L E I N F O

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

A B S T R A C T

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