# A Model-Theoretic Approach to Regular String Relations*

Michael Benedikt† Bell Labs

Leonid Libkin§ U. Toronto

Thomas Schwentick¶ U. Mainz

Luc Segoufin‖ INRIA

## Abstract

*We study algebras of definable string relations – classes of regular $n$-ary relations that arise as the definable sets within a model whose carrier is the set of all strings. We show that the largest such algebra – the collection of regular relations – has some quite undesirable computational and model-theoretic properties. In contrast, we exhibit several definable relation algebras that have much tamer behavior: for example, they admit quantifier elimination, and have finite VC dimension. We show that the properties of a definable relation algebra are not at all determined by the one-dimensional definable sets. We give models whose definable sets are all star-free, but whose binary relations are quite complex, as well as models whose definable sets include all regular sets, but which are much more restricted and tractable than the full algebra of regular relations.*

## 1 Introduction

In the past 40 years, various connections between logic, formal languages and automata have been explored in great detail. The standard setting for connecting logical definability with various properties of formal languages is to represent strings over a finite alphabet $\Sigma = \{a_1, \ldots, a_n\}$ as first-order structures in the signature $(P_{a_1}, \ldots, P_{a_n}, <)$, so that the structure $M_s$ for a string $s$ of length $k$ has the universe $\{1, \ldots, k\}$, with $<$ being the usual ordering, and $P_{a_i}$ being the set of the positions $l$ such that the $l$th character in $s$ is $a_i$. Then a sentence $\Phi$ of some logic $\mathcal{L}$ defines a language $L(\Phi) = \{s \in \Sigma^* \mid M_s \models \Phi\}$. Two classical results on logic and language theory state that languages thus definable in monadic second-order logic (MSO) are precisely the regular languages [8], and the languages definable in first-order logic (FO) are precisely the star-free languages [25]. For a survey, see [28, 29].

An alternative approach to definability of strings, based on classical infinite model theory rather than finite model theory, dates back to [8, 10]. One considers an infinite structure $M$ consisting of $\langle \Sigma^*, \Omega \rangle$, where $\Omega$ is a set of functions, predicates and constants on $\Sigma^*$. One can then look at definable sets, those of the form $\{\vec{a} \mid M \models \varphi(\vec{a})\}$, where $\varphi$ is a first-order formula in the language of $M$. A well-known result links definability with traditional formal language theory. Let $\Omega_{\text{reg}}$ consist of unary functions $l_a$, $a \in \Sigma$, binary predicates $\text{el}(x,y)$ and $x \preceq y$, where $l_a(x) = x \cdot a$, $\text{el}(x,y)$ states that $x$ and $y$ have the same length, and $x \preceq y$ states that $x$ is a prefix of $y$. Let $\mathbf{S}_{\text{len}}$ be the model $\langle \Sigma^*, \Omega_{\text{reg}} \rangle$ (we will explain the notation later). Then subsets of $\Sigma^*$ definable in $\mathbf{S}_{\text{len}}$ are precisely the regular languages [8, 10, 9].

An advantage of the "model-theoretic approach" is that one immediately gets an extension of the notion of recognizability from string languages to $n$-ary string relations for arbitrary $n$. One gets an algebra of $n$-ary string relations for every $n$, and these algebras automatically have closure under projection and product, in addition to the boolean operations. In the case of the model $\mathbf{S}_{\text{len}}$ above, this algebra is not new: in fact, the definable $n$-ary relations are exactly the ones recognizable under a natural notion of automaton running over $n$-tuples [10, 15].

An obvious question to ask, then, is whether *new* algebras of string relations arise through the model-theoretic approach. In particular, if we restrict the signature $\Omega$ to be less expressive than $\Omega_{\text{reg}}$, do we get new relation algebras lying within the recognizable relations?

---

*Part of this work was done while the second and the third authors visited INRIA, and the second and the fourth authors visited Mainz.

†Bell Laboratories, 263 Shuman Blvd, Naperville, IL 60566, USA. E-mail: benedikt@research.bell-labs.com.

§Department of Computer Science, University of Toronto, 6 King's College Road, Toronto, Ontario M5S 3H5, Canada. E-mail: libkin@cs.toronto.edu. Research affiliation: Bell Labs.

¶Johannes Gutenberg-Universität Mainz, Institut für Informatik, 55099 Mainz, Germany. Email: tick@informatik.uni-mainz.de.

‖INRIA-Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France. E-mail: Luc.Segoufin@inria.fr.

A natural starting point would be to find a signature that captures properties of the star-free sets. Here again, a simple example leaps out: consider the signature $\Omega_{\rm sf} = (\preceq, (l_a)_{a\in\Sigma})$, and let $\mathbf{S} = \langle \Sigma^*, \Omega_{\rm sf}\rangle$. One can easily show that the definable subsets of $\Sigma^*$ in $\mathbf{S}$ are exactly the star-free ones. Furthermore, we will show that the definable $n$-ary relations of this model are exactly those definable by regular prefix automata (cf. [1]) whose underlying string automata are counter-free.

Just as there is a significant difference between the complexity-theoretic behavior of regular languages and star-free languages, we find that the model $\mathbf{S}$ is much more tractable, in terms of its model-theory and its complexity than $\mathbf{S}_{\rm len}$. In particular, we show that $\mathbf{S}$ has quantifier-elimination in a natural relational extension, while $\mathbf{S}_{\rm len}$ does not.

It would be tempting to think of $\mathbf{S}$ and $\mathbf{S}_{\rm len}$ as canonical extensions of the notions of regularity and star-free to $n$-ary relations. However, we will show that in fact there are *many* choices for $\Omega$ that share the same one-dimensional definable sets (either star-free or regular). Furthermore, algebras of definable sets may be identical in terms of the string languages they define, but differ considerably in the $n$-ary string relations in the definable algebra. We thus say that an algebra of definable sets based on $\langle \Sigma^*, \Omega\rangle$, with $\Omega \subseteq \Omega_{\rm reg}$ is a *regular algebra of definable sets* if the subsets of $\Sigma^*$ in it (i.e the one-dimensional definable sets of $\langle \Sigma^*, \Omega\rangle$) are exactly the regular sets. We likewise say that the algebra based on definable sets for $\langle \Sigma^*, \Omega\rangle$ is a *star-free algebra of definable sets* if the subsets of $\Sigma^*$ in the algebra are exactly the star-free sets.

The rest of the paper studies new examples of regular and star-free definable algebras. We give an example of a star-free algebra with considerably more expressive power than the basic star-free algebra $\mathbf{S}$. This model, which we denote by $\mathbf{S}_{\rm left}$ (as it allows one to add characters on the *left* of a string), shares most of the desirable properties of $\mathbf{S}$: in particular, it has quantifier-elimination in a natural language, and membership test in this algebra has low complexity.

More surprisingly, perhaps, we give examples of regular algebras (which we denote $\mathbf{S}_{\rm reg}$ and $\mathbf{S}_{\rm reg,left}$) that are strictly contained in $\mathbf{S}_{\rm len} = \langle \Sigma^*, \Omega_{\rm reg}\rangle$. Although the one-dimensional sets in these algebras are still the regular sets, the algebra as a whole shares many of the attractive properties of the star-free languages. In particular, we give quantifier-elimination results for these algebras.

One key motivation for our work comes from the field of databases, in particular, the study of query languages with interpreted operations [3, 5, 19], and constraint databases [23]. In those settings, quantifier-elimination gives one closed-form evaluation for queries; it says that one can evaluate queries whose input is a quantifier-free definable set and get a closed form solution as another quantifier-free definable set. This approach has generally been applied to numerical domains over the reals, since there are several powerful quantifier-elimination results available there. It is natural to extend this approach to databases over strings: the string datatype, after all, is ubiquitous in database applications, and languages such as SQL already give some capability of manipulating star-free sets (via the LIKE predicate) defined from the input data within queries. But in order to extend the constraint-database approach to the string context, we are first required to find definable algebras that admit quantifier-elimination in some natural yet powerful language. (Some of the previous results in this direction considered query languages over undecidable structures [20], or decidable ones but not capable of expressing some very basic operations on strings [14].) The quantifier-elimination results here thus yield new examples where the constraint approach can be applied. In fact, the results we present here were used in [7] to give expressiveness and complexity bounds for the database query languages that arise from several algebras of definable sets.

Our approach was also motivated by the study of *automatic structures* [22, 9], which are a subclass of recursive structures [21], and were introduced recently as a generalization of automatic groups [16]. In an automatic structure $M = \langle \Sigma^*, \Omega\rangle$, every predicate in $\Omega$ is definable by a finite automaton. More precisely, an $n$-ary predicate $P$ is given by a letter-to-letter $n$-automaton [15, 18]. Such an automaton is a usual DFA whose alphabet is $(\Sigma \cup \{\#\})^n$, $\# \notin \Sigma$. An $n$-tuple of strings $s_1, \ldots, s_n$ can be viewed as a word of length $\max_i |s_i|$ over the alphabet $\Sigma \cup \{\#\}$, where the $j$th letter is the tuple $(s_1^j, \ldots, s_n^j)$; here $s_k^j$ is the $j$th letter of $s_k$, if $|s_k| \leq j$, and $\#$ otherwise. We then say that a predicate $P \subseteq (\Sigma^*)^n$ is definable by a letter-to-letter $n$-automaton $A$ if $(s_1, \ldots, s_n) \in P$ iff $A$ accepts $s_1, \ldots, s_n$.

It is known [10, 9] that a structure is automatic iff it can be interpreted in the structure $\mathbf{S}_{\rm len}$; hence $\mathbf{S}_{\rm len}$ is in some sense the universal automatic structure. It is interesting then to look at subclasses of automatic structures definable within $\mathbf{S}_{\rm len}$ that are significantly more restrictive, and that might have stronger model-theoretic or computational properties than a rich structure like $\mathbf{S}_{\rm len}$. One dividing line we focus on is between automatic structures that do admit quantifier-elimination in a natural relational language, and those that do not. Our first result gives a partial answer to open question 0 in [26], which asks whether $\mathbf{S}_{\text{len}}$ itself has quantifier-elimination in a reasonable signature. We show that it does not have quantifier-elimination in any relational signature of bounded arity. The other structures that we study — $\mathbf{S}$, $\mathbf{S}_{\text{reg}}$, $\mathbf{S}_{\text{left}}$ and $\mathbf{S}_{\text{reg,left}}$ — do admit such a quantifier-elimination. A second dichotomy is between automatic structures that admit star-free definable algebras versus those that have regular algebras. We show that the models $\mathbf{S}$ and $\mathbf{S}_{\text{left}}$ have star-free definable algebras, while the model $\mathbf{S}_{\text{reg}}$ does not. Our results indicate that the class of automatic structures that admit star-free definable algebras is richer than one might have guessed.

**Organization** Section 2 introduces the notation. Section 3 explores the motivating example, the model $\mathbf{S}_{\text{len}}$, and proves a set of results concerning its limitations. In Section 4 we turn to the minimal example of a star-free algebra, the model $\mathbf{S}$, and prove a quantifier-elimination result for this model that contrasts with the negative result proved for $\mathbf{S}_{\text{len}}$. Section 5 extends the results of the previous section to a more complex example of a star-free algebra, the model $\mathbf{S}_{\text{left}}$. Section 6 gives a restriction of $\mathbf{S}_{\text{len}}$ that admits a regular algebra, and proves a quantifier elimination result for this model. The section also connects this model to the minimal model $\mathbf{S}$. Section 7 gives an additional example of a regular algebra, which contains each of the previous examples. Section 8 gives conclusions. All proofs are in the full version [6].

## 2 Notations

Throughout the paper, $\Sigma$ denotes a finite alphabet, and $\Sigma^*$ the set of all finite strings over $\Sigma$. We consider a number of operations on $\Sigma^*$:

- $x \cdot y$ – concatenation of two strings $x$ and $y$.
- $x \preceq y$ – $x$ is a prefix of $y$.
- $l_a(x)$, $a \in \Sigma$, is $x \cdot a$ (adds *l*ast character).
- $f_a(x)$, $a \in \Sigma$, is $a \cdot x$ (adds *f*irst character).
- $|x|$ is the length of string $x$.
- $x \sqcap y$ is the longest common prefix of the strings $x$ and $y$.
- $x - y$ – the string $z$ such that $y \cdot z = x$, if it exists, and $\epsilon$ otherwise.

Note that $|x|$ does not return a string, so it is not an operation of $\Sigma^*$. Instead, we often consider the predicate $\text{el}(x, y)$ which is true iff $|x| = |y|$.

We shall consider several structures on $\Sigma^*$. The basic one is the structure $\mathbf{S} = \langle \Sigma^*, \preceq, (l_a)_{a\in\Sigma}\rangle$. We could equivalently use unary predicates $L_a$, where $L_a(x)$ is true for strings of the form $x' \cdot a$. Note that in the presence of $\preceq$, $l_a$ and $L_a$ are interdefinable, and we thus shall use both of them.

We further consider a number of extensions of $\mathbf{S}$. In one of them characters can be added on the *left* as well as on the right. This structure is denoted by $\mathbf{S}_{\text{left}} \stackrel{\text{def}}{=} \langle \Sigma^*, \preceq, (l_a)_{a\in\Sigma}, (f_a)_{a\in\Sigma}\rangle$. Another extension, denoted by $\mathbf{S}_{\text{len}}$, adds *length* comparisons via the el predicate (note that using $\preceq$ and el one can express various relationships between lengths of strings, e.g. $|x|\{=, \neq, <, >\}|y|$, $|x| = |y| + k$ for a constant $k$, etc.). To summarize, we mainly deal with the following structures:

- $\mathbf{S} = \langle \Sigma^*, \preceq, (l_a)_{a\in\Sigma}\rangle$;
- $\mathbf{S}_{\text{left}} = \langle \Sigma^*, \preceq, (l_a)_{a\in\Sigma}, (f_a)_{a\in\Sigma}\rangle$;
- $\mathbf{S}_{\text{len}} = \langle \Sigma^*, \preceq, (l_a)_{a\in\Sigma}, \text{el}\rangle$.

Once we consider regular algebras, we introduce two more structures; however, operations in them will be motivated by quantifier-elimination results for $\mathbf{S}$ and $\mathbf{S}_{\text{left}}$ and thus those structures will be defined later.

There is a very close connection between $\mathbf{S}_{\text{len}}$ and an extension of Presburger arithmetic. Assume that $\Sigma = \{0, 1\}$. Let $\text{val}(n)$, for $n \in \mathbb{N}$, be $n$ in binary, considered as a string in $\Sigma^*$. Let $V_2(n)$ be the largest power of 2 that divides $n$. Then $P \subseteq \mathbb{N}^k$ is definable in $\langle \mathbb{N}, +, V_2\rangle$ iff $\{(\text{val}^{-1}(n_1), \ldots, \text{val}^{-1}(n_k)) \mid (n_1, \ldots, n_k) \in P\}$ is definable in $\mathbf{S}_{\text{len}}$ [8, 10].

**Model theory background** Let $\Omega$ be a finite or countably infinite first-order signature, and $M$ a model over $\Omega$. By $\text{FO}(M)$ we denote the set of all first-order formulae in the language of $\Omega$. The (complete) theory of $M$, $\text{Th}(M)$, is the set of all sentences in $\text{FO}(M)$ true in $M$. Two models $M$ and $M'$ over $\Omega$ are elementary equivalent if $\text{Th}(M) = \text{Th}(M')$.

We say that $M$ admits *quantifier elimination (QE)* if for every formula $\varphi(\vec{x})$ in $\text{FO}(M)$ there is a quantifier-free formula $\varphi'(\vec{x})$ such that $\forall \vec{x}\ \varphi(\vec{x}) \leftrightarrow \varphi'(\vec{x})$ is true in $M$.

For a tuple $\vec{a}$ and a model $M$ over $\Omega$, we let $tp_M(\vec{a})$ be the *type* of $\vec{a}$ in $M$ (the set of all formulae of $\text{FO}(M)$ satisfied by $\vec{a}$), and $atp_M(\vec{a})$ be the atomic type in $M$

(the set of all quantifier-free formulae of FO($M$) satisfied by $\vec{a}$) . If $A$ is a subset of $M$, $tp_M(\vec{a}/A)$ is the type of $\vec{a}$ over $A$ in $M$ (the set of all FO-formulae over $\Omega \cup A$ satisfied by $\vec{a}$).

A *$\omega$-saturated model* $M$ over $\Omega$ is a model such that each consistent type over a finite set $A$ in FO($M$) is satisfied in $M$. It is known [11] that every model $M$ over $\Omega$ has an elementary equivalent $\omega$-saturated model $M^*$.

**Isolation, VC-dimension** Let $T$ be a theory over $\Omega$ and $M$ be a model of $T$. A subset $A$ of $M$ is said to be *pseudo-finite* if $(M, A) \models F(T, P)$, where $P$ is a unary predicate, and $F(T, P)$ is the set of all formulae of FO($\Omega \cup P$) satisfied by all finite sets of elements in any model of $T$.

If $p$ is a type over $A$ in $M$, a subset $q$ of $p$ *isolates* $p$ if $p$ is the only type over $A$ in $M$ containing $q$. A complete theory $T$ over $\Omega$ is said to have the *strong isolation property* if for any model $M$ of $T$ and any pseudo-finite set $A$ and any element $a$ in $M$, there is a finite subset $A_0$ of $A$ such that $tp_M(a/A_0)$ isolates $tp_M(a/A)$. We say that it has the *isolation property* if a countable $A_0$ exists as above.

Isolation is an interesting property in the database context because it implies certain collapse results for query languages [3, 17] and it is used for that purpose in [7]. Here we use it to provide bounds on the VC-dimension of definable families.

For a family $\mathcal{C}$ of subsets of a set $U$, and a set $F \subseteq U$, we say that $\mathcal{C}$ *shatters* $F$ if $\{F \cap C \mid C \in \mathcal{C}\}$ is the powerset of $F$. The *VC-dimension* of $\mathcal{C}$ is the maximum cardinality of a finite set shattered by $\mathcal{C}$ (or $\infty$, if arbitrarily large finite sets are shattered by $\mathcal{C}$). This concept is fundamental to learning theory, as finite VC-dimension of a hypothesis space is equivalent to learnability (PAC-learnability) [2, 4].

Now consider a structure $M = \langle \Sigma^*, \Omega \rangle$, and a FO($M$) formula $\varphi(\vec{x}, \vec{y})$. For each $\vec{a}$, let $\varphi(\vec{a}, M) = \{\vec{b} \mid M \models \varphi(\vec{a}, \vec{b})\}$. The family of sets $\varphi(\vec{a}, M)$, where $\vec{a}$ ranges over all tuples over $M$, is called a *definable family*. We say that $M$ has finite VC-dimension if every definable family has finite VC-dimension. In particular, this implies learnability of concepts defined in FO over $M$.

## 3 Regular algebra based on $\mathbf{S}_{\text{len}}$

As mentioned in the introduction, $\mathbf{S}_{\text{len}}$ is the canonical automatic structure, and relations definable in $\mathbf{S}_{\text{len}}$ are precisely the *regular relations*, that is, $k$-ary definable relations are precisely those given by letter-to-letter $k$-automata [9, 10]. In particular, this gives a normal form for $\mathbf{S}_{\text{len}}$-formulae. We introduce a new type of *length-bounded quantifiers* of the form $\exists|x| \leq |y|$ and $\forall|x| \leq |y|$. A formula $\exists|x| \leq |y|\varphi$ is meant as an abbreviation for $\exists x(|x| \leq |y|) \wedge \varphi$. Since every finite automaton can be simulated by a length-bounded FO($\mathbf{S}_{\text{len}}$) formula, we conclude that each FO($\mathbf{S}_{\text{len}}$) formula is equivalent to a length-bounded FO($\mathbf{S}_{\text{len}}$) formula. Note that this result can also be shown by a straightforward Ehrenfeucht-Fraïssé game argument.

The universal property of $\mathbf{S}_{\text{len}}$ mentioned above indicates that $\mathbf{S}_{\text{len}}$ may be "too rich" in relations for many applications. We present evidence for this by addressing the open question of [12, 26] whether $\mathbf{S}_{\text{len}}$ has quantifier elimination in a reasonable signature. One first needs to define what "reasonable" means here. Clearly, every structure has quantifier elimination in a sufficiently large expansion of the signature: add symbols for all definable predicates, for example. One can thus take reasonable to mean a finite expansion, but this is not satisfactory: for example, Presburger arithmetic has quantifier elimination in an infinite signature $(+, <, 0, 1, (\text{mod } k)_{k>1})$. Note however that in this example, the maximum arity of the predicates and functions is 2. In fact, it appears to be a common phenomenon that when one proves quantifier elimination in an infinite signature, there is an upper bound on the arity of functions and predicates in it.

We thus view this condition as necessary for a signature to be "reasonable". In general, a reasonable signature might contain relation symbols as well as function symbols. Nevertheless, we can rule out the possibility of a reasonable, purely relational signature for which $\mathbf{S}_{\text{len}}$ has quantifier elimination. This is in contrast to the weaker structures that we consider, all of which have quantifier elimination in a relational signature of bounded arity. Let $\mathbf{S}_{\text{len}}^{(n,m)}$ be the expansion of $\mathbf{S}_{\text{len}}$ with all definable predicates of arity at most $n$, and definable functions of arity $m$. We show the following:

**Theorem 1** *(a) For any $n \geq 0$, and $m = 0, 1$, $\mathbf{S}_{\text{len}}^{(n,m)}$ does not have QE. In particular, there is a property definable in $\mathbf{S}_{\text{len}}$ which is not a Boolean combination of at most $n$-ary definable predicates in $\mathbf{S}_{\text{len}}$.*

*(b) $\mathbf{S}_{\text{len}}^{(1,2)}$, the expansion of $\mathbf{S}_{\text{len}}$ with all unary predicates and binary functions, has QE.*

*Proof sketch.* For (a), the property is whether for an $N$-tuple of strings, for sufficiently large $N$, there is a position $i$ such that the $i$th symbol of all $N$ strings is 0. For (b), we show a stronger result, assuming that $\Sigma$ contains $\{0, 1\}$. We prove QE in a signature that contains the bitwise *and*, *or*, and *not* functions, left and right shifts, and the following two functions. $\mathrm{Fil}_\sigma(w)$ has a 1 at position $i$ iff $w[i] = \sigma$ and a 0 otherwise, and $\mathrm{Pat}_{j,k}(w)$ has the same length as $w$ and has a 1 at position $i$ iff $i \bmod k = j$ and a 0 otherwise, where $j < k$.

In cases of both (a) and (b), the proofs are based on automata representations of definable sets, cf. [9]. □

Our next result shows another model-theoretic and computational shortcoming of $\mathbf{S}_{\mathrm{len}}$: namely, a single formula $\varphi(x, y)$ can define a widely varying collection of relations as we let the parameter $x$ vary. We formalize this through the notion of VC-dimension.

**Proposition 1** *There are definable families in* $\mathbf{S}_{\mathrm{len}}$ *that have infinite VC-dimension.* □

## 4 Star-free algebra based on S

We now turn to the most obvious analog of $\mathbf{S}_{\mathrm{len}}$ for the star-free sets. This is the model $\mathbf{S}$, which is the most basic model among those studied in the paper. We show that it has remarkably nice behavior: it admits effective QE in a rather small extension to the signature. This immediately tells us that definable subsets of $\Sigma^*$ are precisely the star-free languages. We then characterize the $n$-dimensional definable relations in $\mathbf{S}$ by their closure properties, and by an automaton model.

Note that $\mathbf{S}$ is very close to strings considered as *term algebras*, that is, to $\langle \Sigma, \epsilon, (l_a)_{a\in\Sigma}\rangle$. It is of course well-known that the theory of arbitrary term algebras is decidable and admits QE [24]. However, adding the prefix relation is not necessarily a trivial addition: for arbitrary term algebras with prefix (subterm), only the *existential* theory is decidable, but the full theory is undecidable [30] (similar results hold for other orderings on terms [13]). The undecidability result of [30] requires at least one binary term constructor; our results indicate that in the simpler case of strings one recovers QE with the prefix relation.

We start with a result that gives a normal form for formulae of $\mathrm{FO}(\mathbf{S})$. Given a set $S$ of strings , we let $\mathrm{Tree}(S)$ be the tree (i.e. the partially-ordered structure) generated by closing $S \cup \{\epsilon\}$ under $\sqcap$. In other words, $\mathrm{Tree}(S)$ is the poset $\langle \{x \sqcap y \mid x, y \in S \cup \{\epsilon\}\}, \prec\rangle$. (Note that for any set of strings $s_1, \ldots, s_k$, there are two indices $i, j \le k$ such that $s_1 \sqcap \ldots \sqcap s_k = s_i \sqcap s_j$.)

A *complete tree-order description* of a vector $\vec{w}$ of variables is the atomic diagram of $\mathrm{Tree}(\vec{w})$ in the language of $\epsilon, \preceq, \sqcap$. In other words, it is a specification of all the $\preceq$ relations that hold and do not hold in $\mathrm{Tree}(\vec{w})$.

For each $L \subseteq \Sigma^*$, let $P_L$ be the set of pairs $(x, y)$ of strings such that $x \preceq y$ and $y - x \in L$. The following lemma is obvious, since it is well-known that star-free sets are first-order definable on string models [25].

**Lemma 1** *For each star free language $L$, there is a formula $\varphi_L(x, y)$ in* $\mathrm{FO}(\mathbf{S})$ *which defines $P_L$.* □

We now give a normal form result for $\mathrm{FO}(\mathbf{S})$.

**Proposition 2** *Every formula $\psi(\vec{x})$ in* $\mathrm{FO}(\mathbf{S})$ *can be effectively transformed into an equivalent formula which is a disjunction of formulae of the form*

$$\gamma(\vec{x}) \wedge \delta(\vec{x})$$

*where $\gamma(\vec{x})$ is a complete tree-order description over $\vec{x}$ and $\delta(\vec{x})$ is a conjunction of formulae of the form $\varphi_L(t(\vec{x}), t'(\vec{x}))$, where $L$ is star-free, $t(\vec{x})$ and $t'(\vec{x})$ are either $\epsilon$ or a term of the form $x_i \sqcap x_j$, and $\gamma(\vec{x})$ implies that $t(\vec{x})$ is an immediate successor of $t'(\vec{x})$ in the tree-order.*

*Proof* is by induction on the structure of $\psi$. □

Let $\mathbf{S}^+$ be the expansion of $\mathbf{S}$ to the signature that contains $\epsilon$, $\sqcap$ and a binary predicate $P_L$ for each star-free language $L$. Note that $\mathbf{S}^+$ is a *definable* expansion of $\mathbf{S}$, as all additional functions and predicates are definable. From the normal form we now immediately obtain:

**Theorem 2** $\mathbf{S}^+$ *admits quantifier elimination.*

*Remark.* As mentioned above there is no need to nest the $\sqcap$-operator. Therefore, $\mathbf{S}^+$ can be turned into a relational signature that admits quantifier elimination as follows. For each star-free $L$ let $P'_L$ be the set of tuples $(s_1, s_2, s_3, s_4)$ of strings for which $P_L(\sqcap(s_1, s_2), \sqcap(s_3, s_4))$. Note, that $\sqcap(s_1, s_2) \preceq \sqcap(s_3, s_4)$ can be expressed as $P_{\Sigma^*}(\sqcap(s_1, s_2), \sqcap(s_3, s_4))$. It is straightforward to check that this signature admits quantifier elimination. In the same way, the quantifier elimination results in the remainder of the paper can be turned into quantifier elimination results in a relational signature.

Note also that $\mathbf{S}^+$ could be considered as an expansion of $\mathbf{S}$ with either functions $l_a$ or predicates $L_a$ in the signature. In the latter case, predicates $L_a$ are not needed as $L_a(x)$ iff $P_{\Sigma^* a}(\epsilon, x)$.

Another corollary of the normal form is that in the language of $\mathbf{S}$, it suffices to use only bounded quantification. That is, we introduce *bounded quantifiers* of the form $\exists x \preceq y$ and $\forall x \preceq y$ (where $\exists x \preceq y\ \varphi$ means $\exists x\ x \preceq y \wedge \varphi$), and let $\mathrm{FO}_b(\mathbf{S})$ be the restriction of $\mathrm{FO}(\mathbf{S})$ to formulae $\varphi(y_1,\ldots,y_k)$ in which all quantifiers are of the form $Qx \preceq y_i$. From the normal form and the fact that each $\varphi_L$ can be defined with bounded quantifiers, we obtain:

**Corollary 1** $\mathrm{FO}_b(\mathbf{S}) = \mathrm{FO}(\mathbf{S})$. □

Finally, we characterize $\mathbf{S}$-definable subsets of $\Sigma^*$ and $(\Sigma^*)^k$. Given a subset $R \subseteq (\Sigma^*)^k$ and a permutation $\pi$ on $\{1,\ldots,k\}$, by $\pi(R)$ we mean the set $\{(s_{\pi(1)},\ldots,s_{\pi(k)}) \mid (s_1,\ldots,s_k) \in R\}$.

**Corollary 2**

*a) A language $L \subseteq \Sigma^*$ is definable in $\mathbf{S}$ iff it is star-free.*

*b) The class of relations definable over $\mathrm{FO}(\mathbf{S})$ is the minimal class containing the empty set, $\{\epsilon\}$, $\{a\}$ $a \in \Sigma$, $\preceq$, $\sqcap$, and closed under Boolean operations, Cartesian product, permutation, and the operation $*$ defined by $L_1 * L_2 = \{(s_1, s_1 \cdot s_2) \mid s_1 \in L_1, s_2 \in L_2\}$ for $L_1, L_2 \subseteq \Sigma^*$.*

*Proof.* a) $\mathbf{S}^+$ formulae in one free variable are Boolean combinations of $P_L(\epsilon, x)$, for $L$ star-free, and thus they define only star-free languages.

b) For one direction notice that $\epsilon$, $\{a\}$, $\prec$, $\sqcap$ are definable in $\mathrm{FO}(\mathbf{S})$, and that $\mathrm{FO}(\mathbf{S})$ is closed under boolean operations, permutation and Cartesian product. The closure under $*$ is an easy consequence of Lemma 1 as $L_1 * L_2$ corresponds to $\{(x,y) \mid \varphi_{L_1}(\epsilon,x) \wedge \varphi_{L_2}(x,y)\}$. The other direction follows from the normal form. □

Note that the projection operation is not needed in the closure result above.

**Automaton** We now give an automaton model characterizing definability in $\mathrm{FO}(\mathbf{S})$. This automaton model corresponds exactly to the counter-free variant of *regular prefix automaton* as defined in [1].

Let us recall the definition of regular prefix automaton. Let $A$ be a finite non-deterministic automaton on strings with state set $Q$, transition relation $\delta$ and initial state $q_0$. We construct from $A$ an automaton $\hat{A} = (\Sigma, Q, q_0, F, \delta)$ accepting $n$-tuples $\vec{w} = (w_1, \cdots, w_n)$ of strings in the following way. $F$ is a subset of $Q^n$ which denotes the accepting states of $\hat{A}$. Let $\mathit{prefix}(\vec{w})$ be the set of all prefixes of all $w_i$. A *run* of $\hat{A}$ over $\vec{w}$ is a mapping $h$ from $\mathit{prefix}(\vec{w})$ to $Q$ which assigns to every node $\alpha \in \mathit{prefix}(\vec{w})$ a state $q \in Q$ such that $h(\epsilon) = q_0$ and, $\beta = l_a(\alpha)$ implies $h(\beta) \in \delta(h(\alpha), a)$. The run is accepting if $(h(w_1), \cdots, h(w_n)) \in F$. The $n$-tuple $\vec{w}$ is accepted by $\hat{A}$ if there is an accepting run of $\hat{A}$ over $\vec{w}$. See [1] for more details.

For each finite non-deterministic automaton $A$ the corresponding automaton $\hat{A}$ is called *regular prefix automaton* (RPA). The subset of $(\Sigma^*)^n$, $n \in \mathbb{N}$, it defines is called a *regular prefix relation* (RPR).

If the automaton $A$ is counter-free then we say that the corresponding automaton $\hat{A}$ is counter-free (CF-PA). The following shows that the relations definable in $\mathrm{FO}(\mathbf{S})$ are exactly those recognizable by a CF-PA.

**Proposition 3** *A relation is definable in* $\mathrm{FO}(\mathbf{S})$ *if and only if it is definable by a counter-free prefix automaton.* □

It should be noted that $\mathrm{FO}(\mathbf{S})$ can also be characterized by means of counter-free deterministic bottom-up automata.

**VC-dimension and Isolation** In addition to quantifier elimination, $\mathbf{S}$ has some further model-theoretic properties that distinguish it from $\mathbf{S}_{\mathrm{len}}$.

**Proposition 4** $\mathrm{Th}(\mathbf{S})$ *has the strong isolation property.* □

As a corollary of the isolation property, we prove that, unlike for $\mathbf{S}_{\mathrm{len}}$, the definable families for $\mathbf{S}$ are learnable. First, we need the following.

**Proposition 5** *Let $M$ be a model with the isolation property. Then its definable families have finite VC-dimension.*

We give two proofs of this result in the full version: one is a complexity-theoretic argument, the other model-theoretic. □

It follows that the model $\mathbf{S}$, unlike $\mathbf{S}_{\mathrm{len}}$, has learnable definable families.

**Corollary 3** *Every definable family in* $\mathbf{S}$ *has finite VC-dimension.* □

## 5 Star-free algebra based on $\mathbf{S}_{\mathrm{left}}$

We now study an example of a star-free algebra, one where the $n$-ary relations in the algebra are more complex than those definable over $\mathbf{S}$. Recall that

$\mathbf{S}_{\text{left}} = \langle \Sigma^*, \preceq, (l_a)_{a\in\Sigma}, (f_a)_{a\in\Sigma}\rangle$; that is, in this structure one can add characters on the left as well as on the right.

Without the prefix relation, this structure was studied in [27], where a quantifier-elimination result was proved, by extending quantifier-elimination for term algebras (in fact [27] showed that term algebras with queues admit QE). However, as in the case of $\mathbf{S}$, which differs from strings as terms algebras in that it has the prefix relation, here, too, the prefix relation complicates things considerably.

We start with an easy observation that $\text{FO}(\mathbf{S}_{\text{left}})$ expresses more relations that $\text{FO}(\mathbf{S})$. Indeed, the graph of $f_a$, $F_a = \{(x, a\cdot x) \mid x \in \Sigma^*\}$ is not expressible in $\text{FO}(\mathbf{S})$, which can be shown by a simple game argument. More precisely, given a number $k$ of rounds, let $n = 2^k+1$ and consider the game on the tuples $(0^n, 10^n)$ and $(0^{n+1}, 10^n)$. By Corollary 1 it is sufficient to play on the prefixes of the participating strings. The duplicator has a trivial winning strategy on the strings $10^n$ and a well-known winning strategy on $0^n$ versus $0^{n+1}$.

Let $\mathbf{S}^+_{\text{left}}$ be the extension of $\mathbf{S}_{\text{left}}$ with the same (definable) functions and predicates we added to $\mathbf{S}^+$ (that is, a constant $\epsilon$ for the empty string, the binary function $\sqcap$ for the longest common prefix, the predicate $P_L(x,y)$ for each star-free language $L$), and the unary function $x \mapsto x - a$, for each $a \in \Sigma$ (which is also definable).

**Theorem 3** *$\mathbf{S}^+_{\text{left}}$ admits quantifier elimination.*

*Proof sketch.* Let $\Omega_{\mathbf{S}^+}$ and $\Omega_{\mathbf{S}^+_{\text{left}}}$ be the first-order signatures of $\mathbf{S}^+$ and $\mathbf{S}^+_{\text{left}}$. Let $M$ be an $\omega$-saturated model over $\Omega_{\mathbf{S}^+_{\text{left}}}$ elementary equivalent to $\mathbf{S}^+_{\text{left}}$. It suffices to prove quantifier elimination in $M$. Note that $M$ can have both finite and infinite strings. To prove QE, we must show that every two tuples of elements of $M$ that have the same *atomic* type, have the same type. Define a *nice term* of $\Omega_{\mathbf{S}^+_{\text{left}}}$ as a term of the form $t(x) = x - a + b$, where $a$ and $b$ are finite strings. Given two tuples $\vec{c}$ and $\vec{d}$ of the same length over $M$, define two relations on them:

- $\vec{c} \equiv \vec{d}$ iff for all sequences $i_1, \ldots, i_k$ from $\{1, \ldots, n\}$ (where $n$ is the length of $\vec{c}$) and all sequences $t_1, \ldots, t_k$ of nice terms:
$$\begin{aligned} & atp_{\mathbf{S}^+}(t_1(c_{i_1}), \ldots, t_k(c_{i_k})) \\ = \ & atp_{\mathbf{S}^+}(t_1(d_{i_1}), \ldots, t_k(d_{i_k})) \end{aligned}$$
- $(c', \vec{c}) \equiv_1 (d', \vec{d})$ iff for all sequences $i_1, \ldots, i_k$ from $\{1, \ldots, n\}$ and all sequences $t_1, \ldots, t_k$ of nice terms:
$$\begin{aligned} & atp_{\mathbf{S}^+}(c', t_1(c_{i_1}), \ldots, t_k(c_{i_k})) \\ = \ & atp_{\mathbf{S}^+}(d', t_1(d_{i_1}), \ldots, t_k(d_{i_k})) \end{aligned}$$

Of course, $(c', \vec{c}) \equiv (d', \vec{d})$ implies $(c', \vec{c}) \equiv_1 (d', \vec{d})$, as the identity is a nice term. We then prove the main lemma, which shows that these two relations coincide; that is, if $(c', \vec{c}) \equiv_1 (d', \vec{d})$, then also $(c', \vec{c}) \equiv (d', \vec{d})$.

Using this, we show that $\equiv$ has the back-and-forth property in $M$ (which is actually stronger than what is needed for quantifier-elimination). The theorem follows from the lemma, as each type of the form $atp_{\mathbf{S}^+}(t_1(c_{i_1}), \ldots, t_k(c_{i_k}))$ is also an atomic type of $\mathbf{S}^+_{\text{left}}$. Hence, the atomic types determine the types. For details, see the full version [6]. □

From the previous theorem we get the following corollaries. First, the back-and-forth property of $\equiv_1$ gives us the following normal form for $\text{FO}(\mathbf{S}^+_{\text{left}})$ formulae.

**Corollary 4** *For every $\text{FO}(\mathbf{S}_{\text{left}})$ formula $\rho(x, \vec{y})$ there is an $\text{FO}(\mathbf{S})$ formula $\rho'(x, \vec{z})$ and a finite set of nice $\mathbf{S}^+_{\text{left}}$ terms $\vec{t}$ such that*

$$\forall x \vec{y}\ \rho(x, \vec{y}) \leftrightarrow \rho'(x, \vec{t}(\vec{y}))$$

*holds in $\mathbf{S}_{\text{left}}$.* □

Then Corollary 4 for the empty tuple $\vec{y}$ and Corollary 2 imply:

**Corollary 5** *Subsets of $\Sigma^*$ definable over $\mathbf{S}_{\text{left}}$ are precisely the star-free languages.* □

For formulae in the language of $\mathbf{S}_{\text{left}}$ (as opposed to $\mathbf{S}^+_{\text{left}}$), we can show that bounded quantification suffices, although the notion of bounded quantification is slightly different here from that used in the previous section. Let $N_p(s)$ be the prefix-closure of $\{s - s_1 + s_2 \mid |s_1|, |s_2| \leq p\}$. Clearly $N_p(s)$ is definable from $s$ over $\mathbf{S}_{\text{left}}$. We then define $\text{FO}_*(\mathbf{S}_{\text{left}})$ as the class of $\text{FO}(\mathbf{S}_{\text{left}})$ formulae $\varphi(\vec{x})$ in which all quantification is of the form $\exists z \in N_p(x_i)$ and $\forall z \in N_p(x_i)$, where $x_i$ is a free variable of $\varphi$ and $p \geq 0$ arbitrary.

**Corollary 6** $\text{FO}_*(\mathbf{S}_{\text{left}}) = \text{FO}(\mathbf{S}_{\text{left}})$. □

**Isolation and VC-dimension** We now show that the results about isolation and VC-dimension extend from $\mathbf{S}$ to $\mathbf{S}_{\text{left}}$.

**Proposition 6** $\text{Th}(\mathbf{S}_{\text{left}})$ *has the isolation property.* □

Since the argument for corollary 3 actually shows that isolation implies finite VC-dimension, we conclude:

**Corollary 7** *Every definable family in $\mathbf{S}_{\text{left}}$ has finite VC-dimension.* □

## 6 Regular algebra extending S

The previous sections presented star-free algebras with attractive properties. We now give an example of a regular algebra that has significantly *less* expressive power than the rich structure $\mathbf{S}_{\text{len}}$, and which shares some of the nicer properties of the star-free algebras in the previous sections.

This algebra can be obtained by considering two possible ways of extending FO(**S**): the first is by adding the predicates $P_L$ for all *regular* languages $L$; that is, predicates $P_L(x,y)$ which hold for $x \preceq y$ such that $y - x \in L$, where $L$ is a regular language. The second extension is by using monadic-second order logic instead of only first-order logic. It turns out that these extensions define exactly the same algebra. We show this, and also show that the resulting regular algebra shares the QE and VC-dimension properties of the star-free algebras defined previously.

Let $\mathbf{S}_{\text{reg}} = \langle \Sigma^*, \preceq, (l_a)_{a\in\Sigma}, (P_L)_{L \text{ regular}}\rangle$. Since it defines arbitrary regular languages in $\Sigma^*$, it is a proper extension of **S**. Every FO($\mathbf{S}_{\text{reg}}$)-definable set is definable over $\mathbf{S}_{\text{len}}$, because the predicates $P_L$ are definable in $\mathbf{S}_{\text{len}}$ (the easiest way to see this is by using the characterization of $\mathbf{S}_{\text{len}}$ definable properties via letter-to-letter automata). Thus, we have:

**Proposition 7** *Subsets of* $\Sigma^*$ *definable over* $\mathbf{S}_{\text{reg}}$ *are precisely the regular languages.* □

Let $\mathbf{S}^+_{\text{reg}}$ be the extension of $\mathbf{S}_{\text{reg}}$ with $\epsilon$ and $\sqcap$. Most of the results about **S** and $\mathbf{S}^+$ from Section 4 can be straightforwardly lifted to $\mathbf{S}_{\text{reg}}$ and $\mathbf{S}^+_{\text{reg}}$. For example, the normal form Proposition 2 holds for $\mathbf{S}_{\text{reg}}$ if one replaces "star-free" with "regular": the proof given in Section 4 applies verbatim. From this normal form we immediately obtain:

**Theorem 4** $\mathbf{S}^+_{\text{reg}}$ *admits quantifier elimination.* □

The normal form result also shows that neither the functions $f_a$ nor the predicate el are definable in $\mathbf{S}_{\text{reg}}$ (the former can also be seen from the fact that $\mathbf{S}_{\text{reg}}$ has QE in a signature of bounded arity, and $\mathbf{S}_{\text{len}}$ does not; for inexpressibility of $f_a$ it suffices to apply the normal form results to pairs of strings of the form $(1 \cdot 0^k, 0^k)$). One can also show, as in the case of **S**, that bounded quantification over prefixes is sufficient.

Our next aim is to show that FO($\mathbf{S}_{\text{reg}}$) gives us exactly the same algebra of definable sets as MSO(**S**).

Notice first that each relation definable in FO($\mathbf{S}_{\text{reg}}$) is already definable in MSO(**S**) because each predicate $P_L$ is definable in MSO. We will show in the following that the converse implication also holds.

The proof relies on a lemma which essentially shows that the monadic second-order type of a tuple of strings only depends on its tree-order type and the monadic second-order types of the paths between the strings and their common prefixes.

For a sequence $\vec{a} = (a_1, \ldots, a_n)$ of strings, let $T_{\vec{a}}$ be the structure $\langle \Sigma^*, \preceq, (L_a)_{a\in\Sigma}, \vec{a}\rangle$.

For each string $w \in \Sigma^*$, let $\mathcal{I}_w$ be the finite structure $\langle I_w, <, (R_a)_{a\in\Sigma}, 1, |w|\rangle$ where $I_w$ is $\{1, \ldots, |w|\}$, $<$ is the usual order and, for each $a \in \Sigma$, $R_a$ is the set of all positions of $w$ that carry the letter $a$. For two strings $u, v \in \Sigma^*$, we write $u \equiv^s_k v$ if $\mathcal{I}_u \equiv_{\text{MSO}_k} \mathcal{I}_v$.

**Lemma 2** *For each* $k > 0$*, there is* $k' > 0$ *such that the following holds. Let* $\vec{a} = (a_1, \ldots, a_n), \vec{b} = (b_1, \ldots, b_n)$ *be sequences of strings for which there is a tree isomorphism* $h : \mathit{Tree}(\vec{a}) \to \mathit{Tree}(\vec{b})$ *such that*

*(i) for each* $i \in \{1, \ldots, n\}$, $h(a_i) = b_i$*, and*

*(ii) whenever* $u$ *is the immediate predecessor of* $v$ *in* $\mathit{Tree}(\vec{a})$ *then* $v - u \equiv^s_k h(v) - h(u)$.

*Then* $T_{\vec{a}} \equiv_{\text{MSO}_k} T_{\vec{b}}$. □

As both conditions (i) and (ii) of the Lemma are expressible in FO($\mathbf{S}_{\text{reg}}$), we obtain:

**Theorem 5** FO($\mathbf{S}_{\text{reg}}$) = MSO(**S**). □

The bounded monadic second-order quantifier $\exists X \preceq y$ is defined as follows. A formula $\exists X \preceq y\ \varphi$ holds if and only if $\exists X(\forall x X(x) \to x \preceq y) \wedge \varphi$ holds. We define $\text{MSO}_b(\mathbf{S})$ by binding all first-order and monadic second-order quantifiers.

From Theorem 5 we can easily derive the following corollaries.

**Corollary 8**

- $\text{MSO}_b(\mathbf{S}) = \text{MSO}(\mathbf{S})$
- *Subsets of* $\Sigma^*$ *definable in* MSO(**S**) *are exactly the regular languages.*

**Automata model, isolation, and VC dimension** It was proved in [1] that Regular Prefix Relations (RPR) (those definable by Regular Prefix Automata (RPA), introduced in Section 4) are exactly those definable in MSO(**S**). Thus Theorem 5 together with the results of [1] gives a new characterization of FO($\mathbf{S}_{\text{reg}}$).

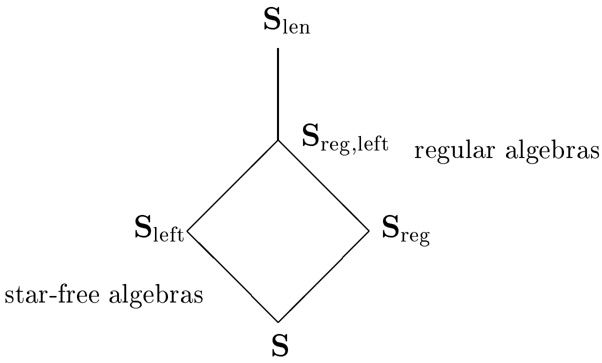


**Figure 1. Relationships between $\mathbf{S}, \mathbf{S}_{\text{left}}, \mathbf{S}_{\text{reg}}, \mathbf{S}_{\text{reg,left}}$, and $\mathbf{S}_{\text{len}}$.**

**Corollary 9** *The relations definable in* $\text{FO}(\mathbf{S}_{\text{reg}})$ *are exactly the RPR relations. Thus each relation definable in* $\text{FO}(\mathbf{S}_{\text{reg}})$ *is recognizable by a RPA.* □

The proof of the isolation property for $\mathbf{S}$ (Proposition 4) is unaffected by the change from star-free $P_L$ to regular $P_L$. Thus, we obtain:

**Corollary 10** $\text{Th}(\mathbf{S}_{\text{reg}})$ *has the isolation property, and definable families of* $\mathbf{S}_{\text{reg}}$ *have finite VC-dimension.* □

## 7 Regular algebra extending $\mathbf{S}_{\text{left}}$

We now give a final example of a regular algebra. Let $\mathbf{S}_{\text{reg,left}}$ be the common expansion of $\mathbf{S}_{\text{left}}$ and $\mathbf{S}_{\text{reg}}$, that is, $\langle \Sigma^*, \preceq, (l_a)_{a\in\Sigma}, (f_a)_{a\in\Sigma}, (P_L)_{L \text{ regular}}\rangle$. Since $\mathbf{S}_{\text{reg}}$ cannot express the functions $f_a$, and $\mathbf{S}_{\text{left}}$ cannot define arbitrary regular sets, we see that $\mathbf{S}_{\text{reg,left}}$ is a proper expansion of $\mathbf{S}_{\text{reg}}$ and $\mathbf{S}_{\text{left}}$. Furthermore, all $\mathbf{S}_{\text{reg,left}}$-definable sets are $\mathbf{S}_{\text{len}}$-definable; the finiteness of VC dimension for $\mathbf{S}_{\text{reg,left}}$, shown below, implies that this containment is proper, too.

Let $\mathbf{S}^+_{\text{reg,left}}$ be the common expansion of $\mathbf{S}^+_{\text{left}}$ and $\mathbf{S}_{\text{reg}}$, that is, the expansion of $\mathbf{S}_{\text{reg,left}}$ with $\epsilon$ and $\sqcap$. The techniques of the previous sections can be used to show the following:

**Theorem 6** $\mathbf{S}^+_{\text{reg,left}}$ *has quantifier-elimination. Furthermore,* $\text{Th}(\mathbf{S}_{\text{reg,left}})$ *has the isolation property, and definable families in* $\mathbf{S}_{\text{reg,left}}$ *have finite VC-dimension.* □

Similarly to $\mathbf{S}_{\text{left}}$, we derive from the proof of Theorem 6 the following normal form for $\mathbf{S}_{\text{reg,left}}$ formulae:

**Corollary 11** *For every* $\text{FO}(\mathbf{S}_{\text{reg,left}})$ *formula* $\rho(x, \vec{y})$ *there is an* $\text{FO}(\mathbf{S}_{\text{reg}})$ *formula* $\rho'(x, \vec{z})$ *and a finite set of nice* $\mathbf{S}^+_{\text{left}}$ *terms* $\vec{t}$ *such that*

$$\forall x\vec{y}\ \rho(x, \vec{y}) \leftrightarrow \rho'(x, \vec{t}(\vec{y}))$$

*holds in* $\mathbf{S}_{\text{reg,left}}$. □

We conclude this section with a remark showing that arithmetic properties definable in structures $\mathbf{S}, \mathbf{S}_{\text{left}}, \mathbf{S}_{\text{reg}}, \mathbf{S}_{\text{reg,left}}$ are weaker than those definable in $\mathbf{S}_{\text{len}}$. As we mentioned earlier, under the binary encoding, $\mathbf{S}_{\text{len}}$ gives us an extension of Presburger arithmetic; namely, it defines $+$ and $V_2$, where $V_2(x)$ is the largest power of 2 that divides $x$. But even $\mathbf{S}_{\text{reg,left}}$ is much weaker:

**Proposition 8** *Neither successor, nor order, nor addition, are definable in* $\mathbf{S}_{\text{reg,left}}$ *(and hence in* $\mathbf{S}, \mathbf{S}_{\text{reg}}, \mathbf{S}_{\text{left}}$*).* □

## 8 Conclusion

There has been significant interest in theoretical computer science in understanding the structure of the regular languages, and in identifying subclasses of the regular languages that have special properties [29, 28]. Our work can be seen as an extension of this program, where we consider subclasses of the regular $n$-ary relations rather than the regular sets. In our approach, however, we do not focus on properties that hold of one particular regular relation by itself, but rather look at some desirable properties of a whole algebra of relations lying within the structure $\mathbf{S}_{\text{len}}$.

We have shown a sharp contrast between the behavior of the full algebra of regular relations of $\mathbf{S}_{\text{len}}$, and those of various submodels such as $\mathbf{S}$, $\mathbf{S}_{\text{left}}$, $\mathbf{S}_{\text{reg}}$, and $\mathbf{S}_{\text{reg,left}}$. We show that the latter are more tractable in many respects. Furthermore, we show that the behavior of an algebra of relations is not at all determined by the one-dimensional sets (subsets of $\Sigma^*$) in the algebra: for example, one can have fairly complex binary relations definable, yet still maintain the property that all definable subsets of $\Sigma^*$ are star-free. Figure 1 summarizes the relationships between the star-free and regular algebras we considered here.

A key question is how many relations one can add to the models $\mathbf{S}_{\text{left}}$ or $\mathbf{S}_{\text{reg}}$ and still have the attractive properties like QE and finite VC-dimension. Is there a model that is somehow maximal with respect to these properties? We would very much like to know the answer to this question. There are also several natural candidate models that would seem amenable to the approach taken here, and where one would expect the same results to go through: for example, if one allows the operation concatenating a fixed sequence "in the middle" of a string, rather than on the left or on the right, is the resulting model still tractable?

## References


[1] D. Angluin, D. N. Hoover. Regular prefix relations. *Mathematical Systems Theory* 17(3),167–191,1984.

[2] M. Anthony and N. Biggs. *Computational Learning Theory.* Cambridge Univ. Press, 1992.

[3] O. Belegradek, A. Stolboushkin, M. Taitslin. Extended order-generic queries. *Annals of Pure and Applied Logic* 97 (1999), 85–125.

[4] A. Blumer, A. Ehrenfeucht, D. Haussler, M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM* 36 (1989), 929–965.

[5] M. Benedikt, L. Libkin. Relational queries over interpreted structures. *J. ACM* 47 (2000), 644–680.

[6] M. Benedikt, L. Libkin, T. Schwentick, L. Segoufin. A model-theoretic approach to regular string relations. INRIA Technical Report, 2000. Available at http://www-rocq.inria.fr/verso/publications/.

[7] M. Benedikt, L. Libkin, T. Schwentick, L. Segoufin. String operations in query languages. In *PODS'01*, pages 183–194.

[8] J.R. Büchi. Weak second-order arithmetic and finite automata. *Zeit. Math. Logik Grundl. Math.* 6 (1960), 66–92.

[9] A. Blumensath and E. Grädel. Automatic structures. In *LICS'00*, pages 51–62.

[10] V. Bruyère, G. Hansel, C. Michaux, R. Villemaire. Logic and $p$-recognizable sets of integers. *Bull. Belg. Math. Soc.* 1 (1994), 191–238.

[11] C.C. Chang and H.J. Keisler Model Theory. North Holland, 1990.

[12] G. Cherlin and F. Point. On extensions of Presburger arithmetic. In *Proc. 4th Easter Model Theory Conf.*, Humboldt Univ. Berlin, 1986.

[13] H. Comon, R. Treinen. The first-order theory of lexicographic path orderings is undecidable. *TCS* 176 (1997), 67–87.

[14] E. Dantsin, A. Voronkov. Expressive power and data complexity of query languages for trees and lists. In *PODS'2000*, pages 157–165.

[15] C. Elgot and J. Mezei. On relations defined by generalized finite automata. *IBM J. Res. Develop.* 9 (1965), 47–68.

[16] D. Epstein et al. *Word Processing in Groups.* Jones and Bartlett Publ., 1992.

[17] J. Flum and M. Ziegler. Pseudo-finite homogeneity and saturation. Preprint, Freiburg University, 1998.

[18] C. Frougny and J. Sakarovitch. Synchronized rational relations of finite and infinite words. *TCS* 108 (1993), 45–82.

[19] E. Grädel and Y. Gurevich. Metafinite model theory. *Information and Computation* 140 (1998), 26–81.

[20] G. Grahne, M. Nykänen, E. Ukkonen. Reasoning about strings in databases. *JCSS* 59 (1999), 116–162.

[21] D. Harel. Towards a theory of recursive structures. In *MFCS'98*, pages 36–53.

[22] B. Khoussainov and A. Nerode. Automatic presentations of structures. In *LCC'94*, pages 367–392.

[23] G. Kuper, L. Libkin, J. Paredaens, editors. *Constraint Databases.* Springer, 2000.

[24] A. Malćev. On the elementary theories of locally free universal algebras. *Soviet Math. Doklady* 2 (1961), 768–771.

[25] R. McNaughton and S. Papert. *Counter-Free Automata.* MIT Press, 1971.

[26] C. Michaux, R. Villemaire. Open questions around Büchi and Presburger arithmetics. In *Logic: From Foundations to Applications*, Oxford Univ. Press, 1996, pages 353–383.

[27] T. Rybina, A. Voronkov. A decision procedure for term algebras with queues. *LICS'2000*, pages 279–290.

[28] H. Straubing. *Finite Automata, Formal Logic, and Circuit Complexity.* Birkhäuser, 1994.

[29] W. Thomas. Languages, automata, and logic. *Handbook of Formal Languages, Vol. 3*, Springer, 1997.

[30] K. Venkataraman. Decidability of the purely existential fragment of the theory of term algebras. *J. ACM* 34 (1987), 492–510.