On Logical Descriptions of Regular Languages

Howard Straubing

Computer Science Department
Boston College
Chestnut Hill, Massachusetts USA 02476
straubin@cs.bc.edu

Abstract. There are many examples in the research literature of families of regular languages defined by purely model-theoretic means (that is, in terms of the kinds of formulas of predicate logic used to define them) that can be characterized algebraically (that is, in terms of the syntactic monoids or syntactic morphisms of the languages). In fact the existence of such algebraic characterizations appears to be the rule. The present paper gives an explanation of the phenomenon: A generalization of Eilenberg's variety theorem is proved, and then applied to logic. We find that a very wide assortment of families of regular languages defined in model-theoretic terms form varieties in this new sense,and that consequently membership in the family depends only on the syntactic morphism of the language.

1 Introduction: Why Logic Leads to Algebra

There is by now an extensive literature on what might be called "descriptive automata theory", in which families of regular languages are classified according to the kinds of logical formulas used to define them. This research began with the work of Büchi on the connection between finite automata and weak second-order arithmetic [4], and continued with results of McNaughton and Papert on first-order definable languages [6], Thomas [14] on the Σ_k-hierarchy, Straubing, Thérien and Thomas [12] on modular quantifiers, and others. Straubing [10] presents a large assortment of such results. More recent work has concentrated on descriptions in temporal logic and the expressive power of formulas with a bounded number of variables (see Thérien and Wilke [13] and Straubing and Thérien [11]).

One intriguing feature of these results is that in almost every case, the family of languages in question can be characterized *algebraically,* in terms of the syntactic monoids or syntactic morphisms of the members the family. It is not particularly obvious why this should be, and the discovery of such an algebraic characterization always comes as something of a pleasant surprise.

The present paper gives a detailed explanation of this phenomenon. Section 2 presents a generalization of the notion of a pseudovariety of finite monoids that includes as special cases both the **M**-varieties and **S**-varieties of Eilenberg's theory [5]. We prove in this setting a generalization of Eilenberg's theorem giving the correspondence between pseudovarieties of monoids and varieties of languages. In Section 3, we apply these ideas to the logical definitions of regular languages. We are able, in this way, to account for the algebraic characterizations of all of the families cited above, and to show that many other such families of languages admit characterizations in terms of the syntactic morphism, even though we do not, at present, possess an explicit description of the underlying class of homomorphisms.

In Section 4 we consider an assortment of related questions concerning locally testable languages, regular languages in circuit complexity classes, and possible extensions of our general theory.

Because of length limitations, we only give a broad outline of our argument. The complete proofs are, for the most part, fairly straightforward, but— especially in Section 3—rather long and repetitive. The full paper will contain all the details.

We assume that the reader is familiar with the fundamentals of the algebraic approach to finite automata, particularly with the notions of the syntactic monoid $M(L)$ and the syntactic morphism μ_L of a regular language L. See Pin [7] for the necessary background.

2 Pseudovarieties of Homomorphisms

2.1 Categories of homomorphisms between free monoids

We consider classes $\mathcal{C}$ of homomorphisms between finitely-generated free monoids with the following properties:
(i) Let Σ, Γ, and Δ be finite alphabets. If $f : \Sigma^* \to \Gamma^*$ and $g : \Gamma^* \to \Delta^*$ are in $\mathcal{C}$, then $g \circ f : \Sigma^* \to \Delta^*$ is in $\mathcal{C}$.
(ii) For each finite alphabet Σ, the identity homomorphism $1_{\Sigma^*} : \Sigma^* \to \Sigma^*$ belongs to $\mathcal{C}$.

Such classes form the morphism classes of *categories* whose objects are all the finitely-generated free monoids. We will abuse terminology slightly by referring to the classes $\mathcal{C}$ themselves as categories. We are most interested in the following three categories:
(i) $\mathcal{C}_{all}$, which consists of all homomorphisms between finitely-generated free monoids.
(ii) $\mathcal{C}_{ne}$, which consists of all *nonerasing* homomorphisms; that is, homomorphisms $f : \Sigma^* \to \Gamma^*$ such that $f(\Sigma) \subseteq \Gamma^+$.
(iii) $\mathcal{C}_{lm}$, which consists of all the *length-multiplying* homomorphisms; that is, homomorphisms $f : \Sigma^* \to \Gamma^*$ for which there exists $k \geq 0$ with $f(\Sigma) \subseteq \Gamma^k$.

2.2 Pseudovarieties of homomorphisms onto finite monoids

Let $\mathcal{C}$ be a category of homomorphisms between free monoids, as described above. A collection $\mathbf{V}$ of homomorphisms $\phi : \Sigma^* \to M$, where Σ is a finite alphabet, M is a finite monoid, and ϕ maps onto M, is a $\mathcal{C}$-*pseudovariety* if the following conditions hold:

(i) Let $\phi : \Sigma^* \to M$ be in $\mathbf{V}$, $f : \Gamma^* \to \Sigma^*$ in $\mathcal{C}$, and suppose there is a homomorphism α from $Im(\phi \circ f)$ onto a finite monoid N. Then $\alpha \circ \phi \circ f : \Gamma^* \to N$ is in $\mathbf{V}$.
(ii) If $\phi : \Sigma^* \to M$ and $\psi : \Sigma^* \to N$ belong to $\mathbf{V}$, then so does $\phi \times \psi : \Sigma^* \to Im(\phi \times \psi) \subseteq M \times N$.
Example. Let $\mathcal{C} = \mathcal{C}_{all}$, and let $\mathbf{V}$ be a $\mathcal{C}$-pseudovariety. Suppose $\phi : \Sigma^* \to M$ is in $\mathcal{C}$, and let $\psi : \Gamma^* \to M$ be a surjective homomorphism. Then there is a homomorphism $f : \Gamma^* \to \Sigma^*$ such that $\psi = \phi \circ f = 1_M \circ \phi \circ f$, and thus ψ is in $\mathbf{V}$. Consequently membership of ϕ in $\mathbf{V}$ depends only on $Im(\phi)$, so we can identify $\mathbf{V}$ with a collection of monoids. It is easy to see that this collection of monoids is closed under direct products and division, and so is a pseudovariety of finite monoids (an "**M**-variety" in Eilenberg's terminology). Conversely, if $\mathbf{V}$ is a pseudovariety of finite monoids, then the collection of all homomorphisms from finitely-generated free monoids onto members of $\mathbf{V}$ forms a $\mathcal{C}_{all}$-pseudovariety.
Example. In a like manner, if $\mathbf{V}$ is a $\mathcal{C}_{ne}$-pseudovariety, then the family of semigroups $\phi(\Sigma^+)$, where $\phi : \Sigma^+ \to S$ is in $\mathbf{V}$, forms a pseudovariety of finite semigroups (an "**S**-variety"). Conversely, if $\mathbf{V}$ is a pseudovariety of finite semigroups, then the family of homomorphisms $\phi : \Sigma^* \to M$, where $\phi(\Sigma^+) \in \mathbf{V}$, is a $\mathcal{C}_{ne}$-pseudovariety.
Example. We show how to generate $\mathcal{C}_{lm}$-pseudovarieties that are not $\mathcal{C}_{ne}$ pseudovarieties. These examples arise in the study of regular languages in circuit complexity classes (see Barrington, *et. al.* [2]); their logical theory is discussed at length in Straubing [10]. Let $\phi : \Sigma^* \to M$ be a homomorphism, with M a finite monoid. The sets $\phi(\Sigma^i) = \phi(\Sigma)^i$, for $i > 0$, form a finite subsemigroup $\mathcal{T}$ of the power semigroup of M, and consequently there is a unique idempotent $S = \phi(\Sigma^k)$ in $\mathcal{T}$. Since $S = S^2$, S is a subsemigroup of M, which we call the *stable subsemigroup* of ϕ. Let $\mathbf{V}$ be a pseudovariety of finite semigroups, and let $\mathbf{W}$ be the family of all surjective homomorphisms $\phi : \Sigma^* \to M$ whose stable subsemigroup belongs to $\mathbf{V}$. Then $\mathbf{W}$ is a $\mathcal{C}_{lm}$-pseudovariety. We omit the proof.

The following lemma, whose simple proof we omit, gives a description of the $\mathcal{C}$-pseudovariety generated by a collection of homomorphisms.

Lemma 1. *Let Φ be a collection of homomorphisms from finitely-generated free monoids onto finite monoids, and let $\mathcal{C}$ be a category of homomorphisms between free monoids. The smallest $\mathcal{C}$-pseudovariety containing Φ consists of all surjective homomorphisms $\alpha : \Sigma^* \to M$ for which there exist finite alphabets $\Sigma_1, \ldots, \Sigma_r$, homomorphisms $\phi_i : \Sigma_i^* \to M_i$ in Φ, and homomorphisms $f_i : \Sigma^* \to \Sigma_i^*$ in $\mathcal{C}$ such that whenever*

$$\phi_i \circ f_i(v) = \phi_i \circ f_i(w)$$

for all $i = 1, \ldots, r$, then $\alpha(v) = \alpha(w)$.

2.3 Varieties of languages and the Eilenberg correspondence

Let $\mathcal{C}$ be a category of homomorphisms between free monoids, and let $\mathbf{V}$ be a $\mathcal{C}$-pseudovariety. Let $\mathcal{V}$ be the mapping that associates to each finite alphabet Σ

the family

$$\mathcal{V}(\Sigma) = \{L \subseteq \Sigma^* : \mu_L \in \mathbf{V}\},$$

where $\mu_L : \Sigma^* \to M(L)$ denotes the syntactic morphism of L. We call $\mathcal{V}$ the *$\mathcal{C}$-variety of languages* corresponding to $\mathbf{V}$.

The two theorems below are due to Eilenberg [5] for the case of **M**- and **S**-varieties, which, as we have seen, are identical to $\mathcal{C}_{all}$- and $\mathcal{C}_{ne}$-pseudovarieties. The generalization given here to arbitrary $\mathcal{C}$-pseudovarieties is entirely straightforward.

Theorem 1. *The mapping $\mathbf{V} \mapsto \mathcal{V}$ is one-to-one.*

Proof. Suppose $\mathbf{V}_1$ and $\mathbf{V}_2$ are $\mathcal{C}$-pseudovarieties of homomorphisms such that

$$\mathbf{V}_1 \mapsto \mathcal{V},$$

and

$$\mathbf{V}_2 \mapsto \mathcal{V}.$$

Let $\phi : \Sigma^* \to M$ be in $\mathbf{V}_1$. As is well known, for each $m \in M$, the syntactic morphism

$$\mu_{\phi^{-1}(m)} : \Sigma^* \to M(\phi^{-1}(m))$$

factors through ϕ, and ϕ in turn factors through

$$\prod_{m\in M} \mu_{\phi^{-1}(m)} : \Sigma^* \to \prod_{m\in M} M(\phi^{-1}(m)).$$

The first of these statements implies that each $\mu_{\phi^{-1}(m)}$ is in $\mathbf{V}_2$, and the second that ϕ is in $\mathbf{V}_2$. Thus $\mathbf{V}_1 \subseteq \mathbf{V}_2$. It follows by symmetry that $\mathbf{V}_1 = \mathbf{V}_2$.

Theorem 2. *Let $\mathcal{V}$ be a mapping that associates to each finite alphabet Σ a family $\mathcal{V}(\Sigma)$ of regular languages in Σ^*. $\mathcal{V}$ is a $\mathcal{C}$-variety of languages if and only if it satisfies the following properties:*
(i) $\mathcal{V}(\Sigma)$ is closed under boolean operations.
(ii) If $L \in \mathcal{V}(\Sigma)$ and $\sigma \in \Sigma$ then the quotients

$$\sigma^{-1}L = \{v \in \Sigma^* : \sigma v \in L\}$$

and

$$L\sigma^{-1} = \{v \in \Sigma^* : v\sigma \in L\}$$

are in $\mathcal{V}(\Sigma)$.
(iii) If $L \in \mathcal{V}(\Sigma)$ and $f : \Gamma^* \to \Sigma^*$ is in $\mathcal{C}$, then $f^{-1}(L) \in \mathcal{V}(\Gamma)$.

Proof. As is well known, the syntactic morphism of $L \subseteq \Sigma^*$ is identical to that of $\Sigma^*\backslash L$, $\mu_{L_1 \cup L_2}$ factors through $\mu_{L_1} \times \mu_{L_2}$, $\mu_{L\sigma^{-1}}$ and $\mu_{\sigma^{-1}L}$ both factor through μ_L, and $\mu_{f^{-1}(L)}$ factors through $\mu_L \circ f$. It follows that any $\mathcal{C}$-pseudovariety of languages satisfies the properties above.

For the converse, let $\mathcal{V}$ satisfy the properties listed in the lemma. Let $\mathbf{V}$ be the smallest $\mathcal{C}$-pseudovariety of homorphisms such that for each finite alphabet Σ, $\mathbf{V}$ contains the syntactic morphisms of all languages in $\mathcal{V}(\Sigma)$. We claim $\mathcal{V}$ is the $\mathcal{C}$-variety of languages corresponding to $\mathbf{V}$. To prove this, it suffices to show that if $L \subseteq \Sigma^*$ and $\mu_L : \Sigma^* \to M(L)$ is in $\mathbf{V}$, then $L \in \mathcal{V}(\Sigma)$. By Lemma 1, if μ_L is in $\mathbf{V}$, then L is a union of sets of the form

$$\bigcap_{i=1}^{r} f_i^{-1} \circ \mu_{L_i}^{-1}(m_i),$$

where each $f_i : \Sigma^* \to \Sigma_i^*$ is in $\mathcal{C}$, each $L_i \subseteq \Sigma_i^*$ is in $\mathcal{V}(\Sigma_i)$, and where the union ranges over a finite set of r-tuples $(m_1, \ldots, m_r) \in M(L_1) \times \cdots \times M(L_r)$. From the definition of the syntactic monoid it follows that each of the sets $\mu_{L_i}^{-1}(m_i)$ is a finite boolean combination of sets of the form $u^{-1}L_i v^{-1}$, for various words u and v over Σ_i. It follows from the stated closure properties of $\mathcal{V}$ that $L \in \mathcal{V}(\Sigma)$.

3 Application to Logic

3.1 Defining regular languages in formal logic

We begin by giving an example of the kinds of logical formulas we are talking about. Consider the sentence

$$\neg\exists x \exists y (y = x + 1 \wedge Q_\sigma x \wedge Q_\sigma y).$$

We interpret this sentence in words over a fixed finite alphabet $\Sigma = \{\sigma, \tau\}$. The variables in the sentence represent positions in the word; *i.e.,* integers in the range from 1 to the length of the word. The formula $Q_\sigma x$ is interpreted to mean "the letter in position x is σ". Thus the sentence above says, "The word does not contain two consecutive occurrences of σ." The sentence thereby defines the language over Σ consisting of all such words.

We also consider sentences containing *modular quantifiers* $\exists^{r \bmod n}$, where $0 \leq r < n$. For example, the sentence

$$\exists x (Q_\sigma x \wedge \exists^{0 \bmod 2} y (y < x))$$

means that the word contains an occurrence of σ in an odd-numbered position (that is, a position x such that the number of positions strictly to the left of x is even).

In general we will consider families of regular languages in which we restrict (i) the kinds of quantifiers that appear in the sentence; (ii) the kinds of *numerical predicates* (*e.g.,* the formulas $y = x + 1$ and $y < x$ in the examples above) that appear in the sentence; (iii) the depth of nesting of the quantifiers; and (iv) the number of variables. A full account of this model-theoretic approach to defining regular languages can be found in [10].

For the class of quantifiers used, we will allow the following three possibilities: (i) ordinary first-order quantifiers; (ii) modular quantifiers with a fixed modulus

n; (iii) both ordinary quantifiers and modular quantifiers of modulus n in combination. For the class of numerical predicates, we allow the following possibilities: (i) $\{=\}$ (equality alone); (ii) $\{=, +1\}$ (successor and equality together; (iii) $\{<\}$ (ordering); (iv) $\{<, +1\}$ (ordering and successor); (v) $\{<, \equiv i \pmod m\}$ (ordering, together with the predicates $x \equiv i \pmod m$ for a fixed modulus m). We make the convention that this last class includes a 0-ary numerical predicate

$$length \equiv i \pmod m$$

which is satisfied by words whose length is congruent to i modulo m. This predicate can be defined by a depth 2 formula using the other predicates in the class, but as we shall see below, we get sharper results if we include it among the atomic formulas. Similarly, the successor relation can be defined in terms of $<$ and thus the class (iv) above might appear superfluous. But this definition requires us to introduce an extra level of quantification and new variables, and would thus affect the statement of our main theorem. On the other hand, $x = y$ can be defined in terms of $<$ (as $\neg((x < y) \vee (y < x))$ without introducing quantifiers or new variables, and thus it really would be superfluous to include equality in the class (iv).

Given one of our permitted choices $\mathcal{Q}$ for the class of quantifiers, $\mathcal{N}$ for the class of numerical predicates, $d \geq 0$ and $r \geq 0$ we define

$$\mathcal{V}_{\mathcal{Q},\mathcal{N},d,r}$$

to be the mapping that associates to each finite alphabet Σ the family

$$\mathcal{V}_{\mathcal{Q},\mathcal{N},d,r}(\Sigma)$$

of languages defined by sentences of quantifier depth no more than d, using quantifiers in $\mathcal{Q}$, numerical predicates in $\mathcal{N}$ and no more than r variables. Note that this notation suppresses mention of the fixed moduli n and m in the quantifier class and the class of numerical predicates.

Our main result is

Theorem 3. *$\mathcal{V}_{\mathcal{Q},\mathcal{N},d,r}$ is a $\mathcal{C}$-pseudovariety, where*
(i) $\mathcal{C} = \mathcal{C}_{all}$ if $\mathcal{N} = \{=\}$ or $\mathcal{N} = \{<\}$.
(ii) $\mathcal{C} = \mathcal{C}_{ne}$ if $\mathcal{N}$ is one of the classes that contains $+1$.
(ii) $\mathcal{C} = \mathcal{C}_{lm}$ if $\mathcal{N} = \{<, x \equiv i \bmod m\}$.

3.2 Sketch of the proof of Theorem 3

Throughout this section we will suppose that the classes $\mathcal{Q}$ of quantifiers and $\mathcal{N}$ of numerical predicates, as well as the alphabet Σ, are fixed. The proof of the main theorem rests on two model-theoretic lemmas, which we shall state shortly; we omit their proofs. Readers familiar with Ehrenfeucht-Fraïssé games will recognize that if only ordinary quantifers were involved, then both lemmas could be proved fairly directly by the application of such games. But game-based arguments are difficult to adapt to modular quantifiers, and so we were obliged, in the full paper, to give a different argument.

In order to prove theorems about sentences (that is, formulas without free variables) safisfied by words, we need to establish results about formulas *with* free variables satisfied by *structures.* Here a structure is a pair (w, I), where $w \in \Sigma^*$ and I is a map from a finite set V of variable symbols into $\{1, \ldots, |w|\}$. A formula whose free variables are contained in $\mathrm{dom}(I)$ can be interpreted in such a structure. The formal semantics are defined in [10]; however, it should be clear from the context what we mean when we say that such a structure satisfies a formula.

Given $\mathcal{Q}$, $\mathcal{N}$, $r \geq 0$ and $d \geq 0$, we define an equivalence relation $\sim_{d,r}$ on structures as follows:

$$(w, I) \sim_{d,r} (w', I')$$

if $\mathrm{dom}(I) = \mathrm{dom}(I')$, and if the two formulas satisfy exactly the same formulas (over the given base of quantifiers and numerical predicates) of depth no more than d and with no more than r variables, whose sets of free variables are contained in $\mathrm{dom}(I)$. Because we fix the alphabet Σ and the moduli allowable in modular quantifiers and numerical predicates, there are only finitely many inequivalent formulas with a given depth and fixed set of free variables. Therefore $\sim_{d,r}$ is an equivalence relation of finite index on the set of structures (w, I) over a given alphabet Σ and finite set of variable symbols $\mathrm{dom}(I)$.

Our fist lemma says that, in a certain sense, $\sim_{d,r}$ is a congruence on structures.

Lemma 2. *Let* $d \geq 0$, $r \geq 0$. *Let* $(w, I) \sim_{d,r} (w', I')$, *and let* $\sigma \in \Sigma$. *Then*

$$(w\sigma, I) \sim_{d,r} (w'\sigma, I'),$$

and

$$(\sigma w, J) \sim_{d,r} (\sigma w', J'),$$

where $\mathrm{dom}(J) = \mathrm{dom}(J') = \mathrm{dom}(I)$, *and for all* $x \in \mathrm{dom}(I)$, $J(x) = I(x) + 1$, $J'(x) = I'(x) + 1$.

Now let (w, I) and (w', I') be $\sim_{d,r}$-equivalent structures, with $w, w' \in \Sigma^*$. Let $f : \Sigma^* \to \Gamma^*$ be a homomorphism. We will define a number of structures over Γ associated with these two structures. We begin with an example: Let $w = \sigma\tau$, $w' = \sigma\tau\tau$, $I(x) = I'(x) = 1$, $I(y) = 2$, and $I'(y) = 3$. The two structures are then $\sim_{0,2}$-equivalent over the base of ordinary quantfiers and $\mathcal{N} = \{<\}$. Let $f(\sigma) = \delta\gamma$ and $f(\tau) = \gamma\delta\gamma$. $f(w)$ then decomposes into the two factors $f(\sigma)f(\tau)$, and $f(w')$ likewise decomposes into three factors. A structure $(f(w), J)$ is *compatible* with (w, I) if the domains of I and J are the same, and if whenever $I(v) = i$, $J(v)$ is a position in the i^{th} factor of $f(w)$. In the present example, $J(x)$ can be 1 or 2, and $J(y)$ can be 3,4 or 5; there are consequently six different such structures compatible with (w, I). Let us pick $J(x) = 2$ and $J(y) = 4$. There is then a unique structure $(f(w'), J')$ that is compatible with (w', I') and *consistent* with $(f(w), J)$ in that if J maps v to the j^{th} position of one of the factors of $f(w)$, then J' maps v to the j^{th} position of the corresponding factor. In our example we must have $J'(x) = 2$ and $J'(y) = 7$.

In general, set

$$w = \sigma_1 \cdots \sigma_k,$$

and

$$w' = \sigma'_1 \cdots \sigma'_{k'}.$$

So

$$f(w) = f(\sigma_1) \cdots f(\sigma_k),$$

and

$$f(w') = f(\sigma'_1) \cdots f(\sigma'_{k'}).$$

If i is a position in $f(w)$, then we set $p(i) = j$ if i belongs to the factor $f(\sigma_j)$, and $q(i) = t$, if i is the t^{th} position in $f(\sigma_j)$. We define analogous mappings p' and q' on the positions of w'. We say $(f(w), J)$ is compatible with (w, I) if I and J have the same domain, and $p(J(v)) = I(v)$ for all v in this domain, and that $(f(w'), J')$ is consistent with $(f(w), J)$ if it is compatible with (w', I'), and if $q(J(v)) = q'(J'(v))$ for all variable symbols v.

Here is our second principal lemma.

Lemma 3. *Let $f : \Sigma^* \to \Gamma^*$ be a $\mathcal{C}$-homomorphism, where*
(i) if $\mathcal{N} = \{=\}$ or $\mathcal{N} = \{<\}$, then $\mathcal{C} = \mathcal{C}_{all}$;
(ii)if $\mathcal{N}$ contains $+1$, then $\mathcal{C} = \mathcal{C}_{ne}$;
(iii) if $\mathcal{N} = \{<, \equiv i \pmod m\}$, then $\mathcal{C} = \mathcal{C}_{lm}$.

Let $d, r \geq 0$ and let (w, I), (w', I') be $\sim_{d,r}$-equivalent structures, with $w, w' \in \Sigma^*$. Let $(f(w), J)$, $f(w'), J')$ be consistent structures compatible with (w, I) and (w', I'), respectively. Then

$$(f(w), J) \sim_{d,r} (f(w'), J').$$

We now complete the proof of Theorem 3. By Theorem 2, it suffices to show that $\mathcal{V}_{Q,\mathcal{N},d,r}$ is closed under boolean operations, quotients, and inverse images of homomorphisms in $\mathcal{C}$. Closure under boolean operations is trivial. To show closure under inverse images, let $L \in \mathcal{V}_{Q,\mathcal{N},d,r}(\Sigma)$, and let $f : \Gamma^* \to \Sigma^*$ be a homomorphism in $\mathcal{C}$. Let $w \in f^{-1}(L)$, and suppose $w \sim_{d,r} w'$, for some $d, r \geq 0$. Then by Lemma 3, $f(w) \sim_d f(w')$, and thus, since L is a union of $\sim_{d,r}$-classes, $f(w') \in L$, so $w' \in f^{-1}(L)$. It follows that $f^{-1}(L)$ is a union of $\sim_{d,r}$-classes, and thus in $\mathcal{V}_{Q,\mathcal{N},d,r}(\Gamma)$.

The identical argument, using Lemma 2, shows closure under quotients.

4 Related Results and Directions for Further Research

4.1 Explicit characterization of logically-defined classes

Which regular languages can be defined by 3-variable first-order sentences of quantifier depth no more than 7 over the base $\{=, +1\}$? We don't know, but we do know, thanks to Theorem 3, that a language belongs to this family if and only if its syntactic semigroup belongs to a particular pseudovariety of finite semigroups—that is, this family of languages forms a $\mathcal{C}_{ne}$-variety of languages. Our work leads to an infinite assortment of such questions, all of which, we now know, have algebraic answers, and some of these answers are likely to be compelling and interesting.

We do, in fact, possess an effective characterization of the pseudovariety of finite semigroups corresponding to the family of languages defined by first-order sentences over $\{=, +1\}$. (See [10].) We conjecture that there is an infinite hierarchy within this family based on the number of variables. (In contrast, if $<$ is included among the numerical predicates, three variables suffice to define all languages in the class.)

For technical reasons, our arguments do not apply to classes defined over the base $\{+1\}$, without the use of equality. In fact, $x = y$ is definable in terms of successor, by the 3-variable formula

$$\forall z(z = x + 1 \leftrightarrow z = y + 1).$$

This leaves open the characterization of the languages definable by 2-variable sentences over this base. In fact it is not difficult to show that this is precisely the class of locally testable languages, which is well known to form a $\mathcal{C}_{ne}$-pseudovariety of languages.(Brzozowski and Simon [3].)

4.2 Universal algebra of pseudovarieties of homomorphisms

There is now a well-developed theory, rooted in universal algebra, concerning pseudovarieties of finite semigroups and monoids. This research, which centers around the use of a special kind of equational description—"pseudoidentities"—to define pseudovarieties, and computations in free profinite semigroups, has become an important tool in the study of finite semigroups. (See Almeida [1].) There is doubtless an extension of this theory to the pseudovarieties of homomorphisms that we study in this paper. As a first step in this direction we pose the problem of giving an equational description—whatever that might mean in this setting—of the $\mathcal{C}_{lm}$-pseudovariety of homomorphisms whose stable semigroups are aperiodic.

We also ask about an extension of our theory to include the *positive* varieties of Pin [8].

4.3 Applications outside of logic

Circuit Complexity. The $\mathcal{C}_{lm}$-pseudovariety of homomorphisms whose stable semigroups are aperiodic first arose in the study of regular languages in circuit complexity classes [2]. It was this work on circuit complexity that to some degree motivated the present paper. Theorem 3 makes it easy to prove that a large assortment of families of regular languages defined by circuits are $\mathcal{C}_{lm}$-varieties of languages, and thus admit characterizations in terms of their syntactic morphisms. For example, let us fix $d > 0$, and consider the regular languages that are boolean combinations of depth d AC^0-languages—that is, the boolean closure of the family of languages recognized by polynomial-size, depth d families of circuits with unbounded fan-in AND and OR gates. It is easy to verify that the resulting family of regular languages satisfies the closure properties of Theorem 2 with $\mathcal{C} = \mathcal{C}_{lm}$, and is consequently a $\mathcal{C}_{lm}$-variety of languages. It would be quite interesting to precisely identify the corresponding pseudovariety of homomorphisms. (By way of comaparison, the union of these families over all d corresponds to the $\mathcal{C}_{lm}$-pseudovariety consisting of all homomorphisms whose stable subsemigroups are aperiodic.)
Generalized Star-Height. We can also consider the category $\mathcal{C}_{lp}$ of *length-preserving* homomorphisms between free monoids. It is a long-standing open problem whether the generalized star-height of regular languages is bounded, or indeed if there are any regular languages whose generalized star-height is strictly greater than 1. See Pin, *et. al.,* [9], where it is proved that the family of languages of generalized star-hight no greater than d, for each fixed d, satisfies the hypotheses of Theorem 2 for $\mathcal{C}_{lp}$. Thus membership of a language in this family is again determined by the syntactic morphism of the language.
Acknowledgments. I wish to thank Jorge Almeida, Jean-Eric Pin, and Denis Thérien for helpful conversations concerning this work.

References

1. J. Almeida, *Finite Semigroups and Universal Algebra*, World Scientific, Singapore, 1994.
2. D. Mix Barrington, K. Compton, H. Straubing, and D. Thérien, "Regular Languages in NC^1", *J. Comp. Syst. Sci.* **44** (1992) 478–499.
3. J. Brzozowski and I. Simon, "Characterizations of Locally Testable Events", *Discrete Math.* **4** (1973) 243–271.
4. J. R. Büchi, "Weak Second-order Arithmetic and Finite Automata",*Z. Math. Logik Grundl. Math.* **6** 66-92 (1960).
5. S. Eilenberg, *Automata, Languages and Machines,* vol. B,Academic Press, New York, 1976.
6. R. McNaughton and S. Papert, *Counter-Free Automata,* MIT Press, Cambridge, Massachusetts, 1971.
7. J. E. Pin, *Varieties of Formal Languages,* Plenum, London, 1986.
8. J.-E. Pin, A variety theorem without complementation, *Russian Mathematics (Izvestija vuzov.Matematika)* **39** (1995), 80–90.

9. J.-E. Pin, H. Straubing and D. Thrien, "New results on the generalized star-height problem", *Information and Computation* **101** 219-250 (1992).
10. H. Straubing, *Finite Automata, Formal Languages, and Circuit Complexity,* Birkhäuser, Boston, 1994.
11. H. Straubing and D. Thérien, "Regular Languages Defined by Generalized First-Order Sentences with a Bounded Number of Bound Variables", in *STACS 2001*, Springer, Berlin 551-562 (2001) (Lecture Notes in Computer Science 2010.)
12. H. Straubing, D. Thérien, and W. Thomas, "Regular Languages Defined by Generalized Quantifiers", *Information and Computation* **118** 289-301 (1995).
13. D. Thérien and T. Wilke, "Over Words, Two Variables are as Powerful as One Quantifier Alternation," *Proc. 30th ACM Symposium on the Theory of Computing* 256-263 (1998).
14. W. Thomas, "Classifying Regular Events in Symbolic Logic", *J. Computer and System Sciences* **25** (1982) 360–376.