Outline of PhD Thesis

http://staff.cs.utu.fi/staff/saeed/pht.html

Author:	Saeed Salehi (Saeed@Math.Net)
Title:	Varieties of Tree Languages
Language:	English
Supervisor:	Magnus Steinby
Reviewers:	Zoltán Ésik Wolfgang Thomas
Opponent:	Thomas Wilke
Institute:	University of Turku, Department of Mathematics; and Turku Centre for Computer Science
Defence Date:	12 August 2005

Abstract

Trees are among the most fundamental and ubiquitous structures in mathematics. Tree languages and automata on trees have been studied extensively since the 1960s from both a purely mathematical and application point of view. When trees are defined as terms, universal algebra becomes directly applicable to tree automata and, on the other hand, the theory of tree automata suggests new notions and problems to universal algebra.

Different syntactic invariants have been proposed as bases for classifications of regular tree languages: syntactic algebras (Steinby 1979, 1992; Almeida 1990), syntactic monoids and syntactic semigroups (Thomas 1983; Nivat and Podelski 1989), tree algebras (Wilke 1996) and syntactic theories (Esik 1999). However, so far variety theorems comparable with Eilenberg's classical theorems for regular string languages were known for syntactic algebras and syntactic theories only. In this thesis we consider several aspects of varieties of tree languages and settle some open questions concerning the various formalisms.

In Chapter 2 we extend the variety theorem for general recognizable subsets of free algebras (Steinby 1979) to the many-sorted case. In Chapter 3 we formulate Pin's (1996) theory of positive varieties for tree languages and prove a variety theorem that establishes a correspondence between positive varieties of tree languages and varieties of finite ordered algebras.

It has been known already for quite a long time that not all varieties of tree languages can be defined by syntactic monoids or semigroups, and the question about the exact defining power of these syntactic invariants has been raised by several authors. In Chapter 4 we answer this question by characterizing the varieties of tree languages that correspond to some variety of finite monoids or semigroups. In Chapter 5 we characterize the families of tree languages definable by ordered monoids and study some special instances of the above mentioned variety theorems.

Chapter 6 is devoted to Wilke's tree algebras. We introduce a convergent term rewriting system that yields an efficient method to decide the word problem of tree algebras. By using the notions introduced in Chapter 2 for many-sorted algebras and languages, we obtain a variety theorem for families of tree languages defined by tree algebras. Moreover, we prove that, for any sufficiently rich alphabet, all congruence-preserving functions of the tree term algebra are obtained as compositions of the basic tree-constructing operations.

Table of Contents

The following is an excerpt of the thesis highlighting its background and new results with more details.

1 Introduction and Preliminaries

Trees and terms are important structured objects that can be found almost everywhere in computer science, not only in connection with their mathematical foundations ([10]). Almost every working mathematician has heard of "trees", as this notion appears in many different areas of mathematics from graph theory to universal algebra to logic. In computer science trees are often regarded as a natural generalization of strings. The theory of tree automata and tree languages emerged in the middle of the 1960s quite naturally from the view of finite automata as unary algebras advocated by J. R. Büchi and J. B. Wright. The theory of tree automata and tree languages can thus be seen as an outgrowth of Büchi's and Wright's program which had as its goal a general theory that would encompass automata, universal algebra, equational logic, and formal languages. Some interesting vistas of this program and its development are opened by Büchi's posthumous book [2] in which many of the ideas are traced back to people like Thue, Skolem, Post, and even Leibniz (see [9]).

Though the theory of tree automata and tree languages may have come into existence by generalizing string automata and languages, but, of course, no branch of mathematics could stay alive very long as a mere generalization. Apart from its intrinsic interest, the theory of tree automata and tree languages has found several applications and it offers new perspectives to various parts of mathematical linguistics. It has also been applied to some decision problems of logic, and it provides tools for syntactic pattern recognition (see [3] and [8]). Actually, using tree automata has proved to be a powerful approach to simplify and extend previously known results, and also to find new results. For instance recent works use tree automata for application in abstract interpretation using set constraints, rewriting, automated theorem proving and program verification, databases and XML schema languages (see [3]).

Mathematicians who have heard of trees may recall one or two definitions of them. Considering trees as terms over a ranked alphabet and a leaf alphabet has become a custom in some schools, especially in Turku, Finland. An advantage of this approach is that the concepts and results to universal algebra become immediately usable. It is worth noting that the impact of universal algebra on the theory of tree automata and tree languages has not been in one direction only; developments of tree automata and tree languages have suggested new problems and concepts to universal algebra. Also in this thesis we have developed algebraic notions and proved theorems in universal algebra when the necessity has emerged. However, the recent book of Denecke and Wismath [4] is the first universal algebra text where tree automata and tree languages are explicitly studied (see its Chapters 5 and 8).

The main topic of the thesis is the variety theory of tree languages and tree automata (finite algebras). The history of variety theory begins with Eilenberg's celebrated variety theorem in [5]. This theorem, which gives a one-to-one correspondence between (pseudo-)varieties of finite semigroups and varieties of recognizable languages, is indeed the most important tool for classifying recognizable languages. Eilenberg's theorem was motivated by characterizations of several families of string languages by syntactic monoids or semigroups (see [5, 14]), above all by Schützenberger's [23] theorem connecting star-free languages and aperiodic monoids. A fascinating feature of this variety theorem is the existence of its many instances. As a matter of fact, most of the interesting classes of algebraic structures are varieties, and similarly, most of the interesting families of tree or string languages studied in the literature turn out to be varieties of some kind. The aforementioned variety theorem connects these interesting families to each other. Eilenberg's theorem has since then been extended in various directions. One of these extensions is Thérien's [27] notion of varieties of congruences on free monoids. Another extension is Pin's positive variety theorem [15] which establishes a bijective correspondence between positive varieties of string languages and varieties of ordered semigroups.

Concerning trees, which are studied in the field of universal algebra, Steinby's variety theorem [24] for varieties of recognizable subsets of free algebras and varieties of finite algebras was the first one of this kind. The correspondence with varieties of congruences, and some other generalizations, were added later by Almeida [1] and Steinby [25, 26]. Another variety theorem for trees is Ésik's [6] correspondence between families of tree languages and classes of theories (see also [7]). As Ésik [6] notes, any variety theorem connects families of tree languages with classes of some structures via their "syntactic structures". One of these syntactic structures is the syntactic semigroup/monoid of a tree language introduced by Thomas [28] and further studied by Salomaa [22]. A different formalism, based on essentially the same concept, was considered by Nivat and Podelski [11, 16].

Several variety theorems for families of recognizable tree languages are proved in this thesis:

- The variety theorem for families of tree languages and varieties of finite algebras, provided by Steinby [24, 25] and Almeida [1], is generalized to many-sorted algebras in Chapter 2 which is a joint work with Steinby [20].

- Chapter 3, based on a joint paper with Petković [13], is inspired by Pin's theory of positive varieties of string languages and varieties of ordered semigroups/monoids. We prove a variety theorem for positive varieties of tree languages and varieties of finite ordered algebras which correspond to each other via syntactic ordered algebras.

- Tree languages definable by syntactic monoids are studied in Chapter 4. It was already known that any family of tree languages definable by syntactic monoids is a (generalized) variety of tree languages, though not every variety of tree languages is definable by syntactic monoids [26]. Characterizing the varieties of tree languages which are definable by syntactic monoids was a relatively long-standing open problem [26, 6]. Here we give an answer to this question by providing a variety theorem for families of tree languages and varieties of finite monoids which correspond to each other via syntactic monoids; the semigroup version of the results is presented as well [19].

- The above characterization of varieties of tree languages definable by semigroups/monoids is generalized to a characterization of positive varieties of tree languages definable by syntactic ordered semigroups/monoids in Chapter 5. This generalization was obtained together with Petković [13]. Also some instances of this positive variety theorem and the variety theorem on Chapter 4 are elaborated.

- The last Chapter 6 is a study of Wilke's tree algebra formalism [30] for binary trees. A completeness theorem for the axiomatization of tree algebras, and a variety theorem for families of binary tree languages and varieties of finite tree algebras is proved in Chapter 6. This variety theorem solves another open problem mentioned by some authors like Ésik [7], Steinby [26], and Wilke [30]. The first two sections of this chapter are based on a joint paper with Steinby [21]. Finally, a completeness property of Wilke's functions is presented without proofs in the last section. We have proved that term algebras over ranked alphabets with at least seven constant symbols are affine-complete (i.e., their every congruence-preserving function is a term function), and also the free tree algebras over finite alphabets containing at least seven labels are affine-complete [17, 18].

The above mentioned results were taken to emphasize the richness of the theory of tree automata and tree languages as they also suggest some new perspectives of the variety theory of string languages when words are viewed as unary trees. This, from a variety theory viewpoint, confirms our belief that not only are trees more than mere generalization of words, but that words are particular cases of trees.

Preliminaries

Strings over a finite alphabet X are often regarded as elements of the free monoid X^* generated by X. Similarly, any tree considered here may be viewed as an element of a term algebra. Also finite tree automata can be defined as finite algebras. Therefore, universal algebra provides a natural mathematical foundation for the theory of finite tree automata and recognizable tree languages. Here we first recall some basic notions of algebras and then present formal definitions and concepts of trees as terms.

A *ranked alphabet* Σ is a finite set of function symbols each of which has a unique non-negative integer arity. For any $m \geq 0$, Σ_m denotes the elements of Σ with arity m. In particular, Σ_0 is the set of constant symbols of Σ. A Σ*-algebra* is a structure $\mathcal{A} = (A, \Sigma)$ where A is a non-empty set in which every symbol of Σ is realized, i.e., any $c \in \Sigma_0$ is realized by a constant $c^{\mathcal{A}} \in A$, and any $f \in \Sigma_m$ for $m > 0$ is realized by an m-ary function $f^{\mathcal{A}} : A^m \to A$. The algebra $\mathcal{A} = (A, \Sigma)$ is called *finite* if the set A is finite.

The notions of *subalgebra, homomorphism, homomorphic image,* and *direct product* of algebras are defined as in universal algebra; see e.g. [29]. A mapping $p : A \to A$ is called an *elementary translation* of $\mathcal{A}$, if for some $m > 0$, $f \in \Sigma_m$, $1 \leq i \leq m$, and $a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_m \in A$, $p(\xi) = f^{\mathcal{A}}(a_1, \ldots, a_{i-1}, \xi, a_{i+1}, \ldots, a_m)$, where ξ is a new variable ranging over A. The set $\mathrm{Tr}(\mathcal{A})$ of *translations* of $\mathcal{A}$ is the smallest set of unary operations on A that contains the *identity map* $1_A : A \to A$, $a \mapsto a$, and all the elementary translations of $\mathcal{A}$, and is closed under the composition.

We call a class of finite Σ-algebras a *variety of finite algebras*, called also pseudo-variety by some authors, if it is closed under subalgebras, homomorphic images and finite direct products of algebras. It can be easily seen that the intersection of any class of varieties is a variety. So, for a collection of Σ-algebras $\mathbf{C}$, the intersection of all varieties which contain $\mathbf{C}$ is a variety, the *variety generated by* $\mathbf{C}$.

Now we define trees as terms. Roughly speaking, a tree is a structured object that is branched from a root which stands in the highest level and every node in the middle is either branched to other nodes or stands as a leaf. For a formal definition, let Σ be a ranked alphabet and X be any finite alphabet, called a *leaf alphabet.* We usually assume $X \cap \Sigma = \emptyset$. The set of ΣX*-trees*, denoted by $\mathrm{T}(\Sigma, X)$, is defined to be the smallest set containing $\Sigma_0 \cup X$ such that $f(t_1, \ldots, t_m) \in \mathrm{T}(\Sigma, X)$ whenever $f \in \Sigma_m$ $(m > 0)$ and $t_1, \ldots, t_m \in \mathrm{T}(\Sigma, X)$. In this formalism the leaves of ΣX-trees are labelled by symbols from $\Sigma_0 \cup X$ and the inner nodes are labelled by the symbols in Σ with non-zero arities. Any subset of $\mathrm{T}(\Sigma, X)$ is called a *tree language.*

The ΣX*-term algebra* $\mathcal{T}(\Sigma, X) = (\mathrm{T}(\Sigma, X), \Sigma)$ is defined by $c^{\mathcal{T}(\Sigma,X)} = c$ for any $c \in \Sigma_0$ and $f^{\mathcal{T}(\Sigma,X)}(t_1, \ldots, t_m) = f(t_1, \ldots, t_m)$ for all $f \in \Sigma_m$ $(m > 0)$ and $t_1, \ldots, t_m \in \mathrm{T}(\Sigma, X)$. We note that $\mathcal{T}(\Sigma, X)$ is the free Σ-algebra generated by X, i.e., for any algebra $\mathcal{A} = (A, \Sigma)$, any mapping $\alpha : X \to A$ can uniquely be extended to a homomorphism $\alpha^{\mathcal{A}} : \mathcal{T}(\Sigma, X) \to \mathcal{A}$. A tree language $T \subseteq \mathrm{T}(\Sigma, X)$ is said to be *recognized* by an algebra $\mathcal{A} = (A, \Sigma)$ when there exist a mapping $\alpha : X \to A$ and a subset $F \subseteq A$ such that $T = F(\alpha^{\mathcal{A}})^{-1} = \{t \in \mathrm{T}(\Sigma, X) \mid t\alpha^{\mathcal{A}} \in F\}$.

Let ξ be a new symbol which does not appear in any ranked alphabet or leaf alphabet considered here. The set $\mathrm{C}(\Sigma, X)$ of ΣX*-contexts* consists of the $\Sigma(X \cup \{\xi\})$-trees in which ξ appears exactly once. For any contexts $P, Q \in \mathrm{C}(\Sigma, X)$ and any tree $t \in \mathrm{T}(\Sigma, X)$, the context $P(Q)$, the *composition* of P and Q, results from P by replacing the special leaf ξ with Q, and the term $P(t)$ results from P by replacing ξ with t. We note that $\mathrm{C}(\Sigma, X)$ is a monoid with the composition operation and identity element ξ. For a tree language $T \subseteq \mathrm{T}(\Sigma, X)$ and context $P \in \mathrm{C}(\Sigma, X)$, the *inverse translation* of T under P is, by definition, the set $P^{-1}(T) = \{t \in \mathrm{T}(\Sigma, X) \mid P(t) \in T\}$.

We shall now give a brief overview of the basic theory of varieties of recognizable tree languages that is the general starting point of this work. Fix Σ be a ranked alphabet. For a tree language $T \subseteq \mathrm{T}(\Sigma, X)$, the *syntactic congruence* $\approx^T$ of T [25] is defined by the following for $t, s \in \mathrm{T}(\Sigma, X)$:

$$t \approx^T s \iff \forall P \in \mathrm{C}(\Sigma, X)\big(P(t) \in T \leftrightarrow P(s) \in T\big).$$

The relation $\approx^T$ is indeed a congruence on $\mathcal{T}(\Sigma, X)$. The *syntactic algebra* $\mathrm{SA}(T)$ of T is the quotient Σ-algebra $\mathcal{T}(\Sigma, X)/\approx^T$. A tree language is recognizable iff its syntactic algebra is finite. It can be shown that $\mathrm{SA}(T)$ is the smallest algebra which recognizes T.

A *family* $\mathscr{V} = \{\mathscr{V}(\Sigma, X)\}$ of recognizable Σ-tree languages is a mapping which assigns to every leaf alphabet X a collection $\mathscr{V}(\Sigma, X)$ of recognizable ΣX-tree languages. A *variety of tree languages* is a family of recognizable tree languages closed under finite Boolean operations (complements, finite unions, and finite intersections), inverse translations and inverse morphisms. That is to say, a family $\mathscr{V} = \{\mathscr{V}(\Sigma, X)\}$ is a variety of tree languages, if for any leaf alphabets X, Y, tree languages $T, T' \subseteq \mathrm{T}(\Sigma, X)$, homomorphism $\varphi : \mathcal{T}(\Sigma, Y) \to \mathcal{T}(\Sigma, X)$, and context $P \in \mathrm{C}(\Sigma, X)$, if $T, T' \in \mathscr{V}(\Sigma, X)$, then $\mathrm{T}(\Sigma, X) \setminus T, T \cap T', T \cup T', P^{-1}(T) \in \mathscr{V}(\Sigma, X)$ and $T\varphi^{-1} \in \mathscr{V}(\Sigma, Y)$. Likewise, a *generalized family* $\mathscr{W} = \{\mathscr{W}(\Sigma, X)\}$ of recognizable tree languages is a mapping which assigns to every pair (Σ, X), where Σ ranges over all finite ranked alphabets and X ranges over all leaf alphabets, a collection $\mathscr{W}(\Sigma, X)$ of recognizable ΣX-tree languages. In [26] Steinby develops a generalized variety theory by introducing the generalized notions of subalgebra, homomorphisms, and direct products. Thus, *generalized variety of tree languages* is a generalized family of recognizable tree languages closed under finite Boolean operations, inverse translations and inverse generalized morphisms.

The á la Eilenberg variety theorem for trees states that the complete lattice of varieties of finite algebras (under the inclusion relation) and the complete lattice of varieties of recognizable tree languages are isomorphic. The isomorphism maps a variety of finite algebras $\mathbf{K}$ to the family of tree languages whose syntactic algebras belong to $\mathbf{K}$ (which can be shown to be a variety of recognizable tree languages), and its inverse maps any variety of recognizable tree languages $\mathscr{V}$ to the class of algebras that can recognize only the tree languages in $\mathscr{V}$ (which in turn can be shown to be a variety of finite algebras).

2 Many-sorted variety theorem

Many-sorted algebras have found their way into computer science through abstract data type specifications. Many-sorted algebras and their specifications in terms of equations or other axioms are the mathematical fundament of rigorous approaches to abstract data types in programming and specification languages. Below we briefly review some definitions.

Fix S to be a set of sorts. An *S-sorted* set $A = \langle A_s \rangle_{s \in S}$ is an S-indexed family of of sets; for each $s \in S$, A_s is the set of elements of sort s in A. Treating S as an alphabet, S^* denotes the set of finite strings over S, including the empty string e, and S^+ is the set of non-empty strings over S. An S-sorted *signature* Ω is a set of operation symbols each of which has been assigned a type that is an element of $S^* \times S$. For any $(w, s) \in S^* \times S$, let $\Omega_{w,s}$ be the set of symbols of type (w, s), and Ω may be given by specifying the non-empty sets $\Omega_{w,s}$. If $f \in \Omega_{w,s}$, then w is the *domain type* of f, and s is its *sort.* In particular, every element of $\Omega_{e,s}$, for the empty string e, is a *constant symbol* of sort s.

An *Ω-algebra* $\mathcal{A} = (A, \Omega)$ consists of an S-sorted set $A = \langle A_s \rangle_{s \in S}$, where $A_s \neq \emptyset$ for every $s \in S$, equipped with constants and operations as follows:
(1) for each constant symbol $c \in \Omega_{e,s}$ of sort $s \in S$, an element $c^{\mathcal{A}} \in A_s$ is specified;
(2) for any function symbol $f \in \Omega_{w,s}$ with $w \in S^+$ and $s \in S$, an operation $f^{\mathcal{A}} \colon A^w \to A_s$ of type (w, s) is assigned. Here $A^w = A_{s_1} \times \cdots \times A_{s_m}$ for $w = s_1 \cdots s_m$.

For many-sorted algebras, the notions of subalgebra, homomorphism and direct product are defined in a sort-wise manner [29]. A variety of many-sorted algebras is a class of finite Ω-algebras (for a fixed Ω) which is closed under subalgebras, homomorphic images and finite direct products. For an S-sorted leaf alphabet $X = \langle X_s \rangle_{s \in S}$, the set of ($S$-sorted) ΩX-trees is defined in a straightforward way.

In Chapter 2 Magnus Steinby and I have considered varieties of recognizable subsets of many-sorted finitely generated free algebras over a given variety and varieties of finite many-sorted algebras [20]. A variety theorem that establishes a bijection between these classes of varieties is proved. For this, appropriate notions of many-sorted syntactic congruences and algebras are needed. Indeed, by developing a theory of varieties of recognizable subsets of free many-sorted algebras we have generalized the theories of [24, 25] and [1] to the many-sorted case. One crucial stage in this development was a thorough analysis of translations in many-sorted algebras which required special care in dealing with different sorts.

3 Positive varieties of tree languages

There are some interesting families of tree languages that do not possess all of the closure properties of varieties. Some of those families are so-called *positive varieties* of tree languages which are families of tree languages closed under finite positive Boolean operations (intersections and unions), inverse translations and inverse morphisms. One example is the family of finite tree languages (note that this family is not closed under complements). These families can not be characterized by algebras, but there is a characterization for them by richer structures, namely by ordered algebras. The theory of ordered algebras is a useful and interesting area in itself, and indeed ordered algebras play an important role in theoretical computer science.

An *ordered* Σ*-algebra* is a structure $\mathcal{A} = (A, \Sigma, \leqslant)$ where (A, Σ) is an algebra and $\leqslant$ is an order on A compatible with the operations of $\mathcal{A}$; that is to say, for any function symbol $f \in \Sigma_m$ $(m > 0)$ and any $a_1, \ldots, a_m, b_1, \ldots, b_m \in A$, whenever $a_1 \leqslant b_1, \ldots, a_m \leqslant b_m$ then $f^{\mathcal{A}}(a_1, \ldots, a_m) \leqslant f^{\mathcal{A}}(b_1, \ldots, b_m)$. We note that any algebra (A, Σ) in the classical sense is an ordered algebra $(A, \Sigma, =)$ in which the order relation is equality. An *order subalgebra* of $\mathcal{A} = (A, \Sigma, \leqslant)$ is a structure $\mathcal{B} = (B, \Sigma, \leqslant')$ such that (B, Σ) is a subalgebra of (A, Σ) and $\leqslant'$ is the restriction of $\leqslant$ to B. An *order morphism* between ordered algebras is an algebraic homomorphism which preserves the orders. Finally, *direct products* of ordered algebras can be defined in a straightforward way. A *variety of ordered algebras* is a class of ordered Σ-algebras (for a fixed Σ) that is closed under order subalgebras, order homomorphic images, and direct products. An *ideal* of an ordered algebra $\mathcal{A} = (A, \Sigma, \leqslant)$ is a downward closed subset $I \subseteq A$; i.e., if $a \leqslant b \in I$ then $a \in I$. A tree language $T \subseteq T(\Sigma, X)$ is *recognized* by the ordered algebra $\mathcal{A}$, if for some ideal I and some morphism $\varphi : \mathcal{T}(\Sigma, X) \to \mathcal{A}$ we have $T = I\varphi^{-1}$. One can prove that for a variety of ordered algebras $\mathbf{K}$, the family of all tree languages recognized by members of $\mathbf{K}$ is a positive variety of tree languages.

The *syntactic quasi-order* $\preccurlyeq_T$ of a tree language $T \subseteq T(\Sigma, X)$ is defined by the following: for any $t, s \in \mathrm{T}(\Sigma, X)$, $t \preccurlyeq_T s \iff (\forall P \in \mathrm{C}(\Sigma, X))\big(P(s) \in T \Rightarrow P(t) \in T\big)$. It induces the following order on the syntactic algebra $(\mathcal{T}(\Sigma, X)/{\approx_T}, \Sigma)$ of T: $t/{\approx^T} \leqslant_T s/{\approx^T} \iff t \preccurlyeq_T s$. The structure $(\mathcal{T}(\Sigma, X)/{\approx_T}, \Sigma, \leqslant_T)$ is indeed an ordered algebra, called the *syntactic ordered algebra* of T. This is the minimal ordered algebra which recognizes T. In [13] Tatjana Petković and I have proved a variety theorem for positive varieties of tree languages and varieties of ordered algebras (Chapter 3). This result is inspired by Pin's positive variety theorem [15] which established a bijective correspondence between positive varieties of string languages and varieties of ordered semigroups/monoids. We also have extended the positive variety theorem to generalized varieties.

4 Definability by monoids

Syntactic monoids of tree languages were introduced by Thomas [28], and further studied by Salomaa [22] and by Nivat and Podelski [11], as a useful structure for studying recognizable tree languages. For example the variety of aperiodic tree languages were characterized by aperiodic monoids in [28].

In Chapter 4 a variety theorem that establishes a bijective correspondence between general varieties of tree languages definable by syntactic monoids and varieties of finite monoids is proved (see also [19]). This solves a problem which has been open for about a decade [6, 26]. It was already known that any family of tree languages definable by syntactic monoids is a (generalized) variety of tree languages [26], though not every variety of tree languages is definable by syntactic monoids; one example is the family of reverse definite tree languages, cf. [30]. In the same chapter, the classes of algebras definable by (translation) monoids are characterized.

Let $\mathcal{A}$ be a finite algebra. With each translation p in $\mathrm{Tr}(\mathcal{A})$ we associate a unary function symbol $\overline{p}$. Let $\Lambda_{\mathcal{A}} = \{\overline{p} \mid p \in \mathrm{Tr}(\mathcal{A})\}$ be the unary ranked alphabet formed by these symbols and let the $\Lambda_{\mathcal{A}}$-algebra $\mathcal{A}^{\varrho} = (\mathrm{Tr}(\mathcal{A}), \Lambda_{\mathcal{A}})$ be defined by $\overline{p}^{\mathcal{A}^{\varrho}}(q) = p(q)$ for all $p, q \in \mathrm{Tr}(\mathcal{A})$. In Proposition 4.1.7, I showed that a class $\mathbf{K}$ of finite algebras is definable by translation monoids if and only if $\mathbf{K}$ is a generalized variety of finite algebras and $\mathcal{A} \in \mathbf{K} \iff \mathcal{A}^{\varrho} \in \mathbf{K}$ holds for any $\mathcal{A}$.

For stating the variety theorem of tree languages and monoids, we need some definitions.

Let $T \subseteq T(\Sigma, X)$ be a tree language. The *syntactic monoid congruence* $\sim^T$ of T on the monoid $\mathrm{C}(\Sigma, X)$ is defined by the following for $P, Q \in \mathrm{C}(\Sigma, X)$,

$$P \sim^T Q \iff \forall R \in \mathrm{C}(\Sigma, X) \forall t \in \mathrm{T}(\Sigma, X) \big(R(P(t)) \in T \leftrightarrow R(Q(t)) \in T \big);$$

and the *syntactic monoid* $\mathrm{SM}(T)$ of T is the quotient monoid $\mathrm{C}(\Sigma, X)/{\sim^T}$. Salomaa proved in [22] that $\mathrm{SM}(T) \cong \mathrm{Tr}(\mathrm{SA}(T))$ holds for any tree language T.

Let Σ and Ω be ranked alphabets, and X and Y be leaf alphabets. A *tree homomorphism* is a mapping $\varphi : \mathrm{T}(\Sigma, X) \to \mathrm{T}(\Omega, Y)$ determined by a mapping $\varphi_X : X \to \mathrm{T}(\Omega, Y)$, and by some mappings, for any $m \geq 0$ with $\Sigma_m \neq \emptyset$, $\varphi_m : \Sigma_m \to \mathrm{T}(\Omega, Y \cup \{\xi_1, \ldots, \xi_m\})$, where the ξ_i's are new variables, inductively as:
(1) $x\varphi = \varphi_X(x)$ for $x \in X$, $c\varphi = \varphi_0(c)$ for $c \in \Sigma_0$, and
(2) $f(t_1, \ldots, t_n)\varphi = \varphi_n(f)[\xi_1 \leftarrow t_1\varphi, \ldots, \xi_n \leftarrow t_n\varphi]$ for each $f \in \Sigma_n$ $(n \geq 1)$ and any $t_1, \ldots, t_n \in \mathrm{T}(\Sigma, X)$, where each ξ_i is replaced with $t_i\varphi$ (for any $i = 1, \ldots, m$); cf. [26].

A tree homomorphism $\varphi : \mathrm{T}(\Sigma, X) \to \mathrm{T}(\Omega, Y)$ is called *regular* if for every $f \in \Sigma_m$ $(m \geq 1)$, each $\xi_1, \ldots, \xi_m$ appears exactly once in $\varphi_m(f)$. The unique extension $\varphi_* : \mathrm{C}(\Sigma, X) \to \mathrm{C}(\Omega, Y)$ of a regular tree homomorphism φ to contexts is obtained by setting $\varphi_*(\xi) = \xi$ (cf. [26], Proposition 10.3). We note that $P(Q)\varphi_* = P\varphi_*(Q\varphi_*)$ and $P(t)\varphi = P\varphi_*(t\varphi)$ hold for all $P, Q \in \mathrm{C}(\Sigma, X)$, $t \in \mathrm{T}(\Sigma, X)$. A regular tree homomorphism $\varphi : \mathrm{T}(\Sigma, X) \to \mathrm{T}(\Omega, Y)$ is said to be *full with respect to* a tree language $T \subseteq \mathrm{T}(\Omega, Y)$, if for every context $Q \in \mathrm{C}(\Omega, Y)$ and every tree $s \in \mathrm{T}(\Omega, Y)$, there are $P \in \mathrm{C}(\Sigma, X)$ and $t \in \mathrm{T}(\Sigma, X)$, such that $Q \sim^T P\varphi_*$ and $s \approx^T t\varphi$. It can be shown that for any regular tree morphism φ and any tree language T, $\mathrm{SM}(T\varphi^{-1})$ is a homomorphic image of a sub-monoid of $\mathrm{SM}(T)$, and when φ is full with respect to T, the monoids $\mathrm{SM}(T\varphi^{-1})$ and $\mathrm{SM}(T)$ are isomorphic.

Now, the variety theorem (Proposition 4.2.14) for tree languages and monoids [19] is as follows.

A family of recognizable tree languages $\mathscr{V} = \{\mathscr{V}(\Sigma, X)\}$ is definable by syntactic monoids if and only if $\mathscr{V}$ is a generalized variety of tree languages and moreover satisfies the following:

(M1) For any regular tree homomorphism $\varphi : \mathrm{T}(\Sigma, X) \to \mathrm{T}(\Omega, Y)$ and any tree language $T \subseteq T(\Omega, Y)$, $T \in \mathscr{V}(\Omega, Y) \Longrightarrow T\varphi^{-1} \in \mathscr{V}(\Sigma, X)$;

(M2) For any regular tree homomorphism $\varphi : \mathrm{T}(\Sigma, X) \to \mathrm{T}(\Omega, Y)$ which is full with respect to a tree language $T \subseteq \mathrm{T}(\Omega, Y)$, $T\varphi^{-1} \in \mathscr{V}(\Sigma, X) \Longrightarrow T \in \mathscr{V}(\Omega, Y)$;

(M3) For every unary ranked alphabet Λ (i.e., $\Lambda = \Lambda_1$), and any leaf alphabets X and Y, we have $Y \subseteq X \Longrightarrow \mathscr{V}(\Lambda, Y) \subseteq \mathscr{V}(\Lambda, X)$.

After proving this variety theorem, it turned out that the result of the well-cited paper [12] (published in 1989) is not correct. That is to say, the family of definite tree languages is not definable by syntactic monoids or semigroups. A concrete example showing that the "if" part of Theorem 1 of [12] does not necessarily hold, can be constructed by considering the tree language over a two-letter alphabet, say $\{\mathsf{a}, \mathsf{b}\}$, which consists of binary trees whose leftmost leaves are labelled with a. One reason for the error in [12] is the misconception that the syntactic semigroup is the syntactic monoid minus the identity element; the authors overlooked the possibility that the identity element can be obtained as the product of non-identity elements (see Example 4.2.2, Remark 4.2.3, and Example 4.3.5 in the thesis).

5 Definability by ordered monoids

Chapter 5 contains the ordered versions of the results of Chapter 4. In particular, it was shown (Proposition 5.2.11) that a (generalized) family of recognizable tree languages is definable by ordered monoids iff it is a (generalized) positive variety that satisfies the above conditions (M1),(M2),(M3). So, it seems that dropping the closure under the complements on the language side is equivalent to considering the ordered version of the (syntactic) structure side (cf. positive variety theorem of [6]).

Also, some interesting families of tree languages have been characterized by algebras and ordered algebras, which provide some natural instances for the variety theorems of Chapters 4 and 5.

6 Tree algebras

A further syntactic structure for recognizable tree languages is introduced by Wilke [30]. In this formalism one considers only binary trees that are represented by terms over a signature Γ consisting of six operation symbols involving the three sorts **label**, **tree** and **context**. A tree algebra is a Γ-algebra satisfying certain identities which identify some pairs of Γ-terms that represent the same tree. The syntactic tree algebra of a tree language T is defined in a natural way. Its component of sort **tree** is the syntactic algebra of T while its **context**-component is the syntactic semigroup of T.

Such binary trees can also be defined as terms: with every $a \in A$ we associate a constant symbol c_a and a binary function symbol f_a. The ranked alphabet $\Sigma^A = \Sigma^A_0 \cup \Sigma^A_2$ is associated with A, where $\Sigma^A_0 = \{c_a \mid a \in A\}$ and $\Sigma^A_2 = \{f_a \mid a \in A\}$.

The sets T_A and C_A of *A-trees* and *A-contexts*, respectively, are defined inductively by:

(1) $c_a \in T_A$ for all $a \in A$, and $\xi \in C_A$;

(2) $f_a(s,t) \in T_A$ and $f_a(p,t), f_a(t,p) \in C_A$ for all $a \in A$, $s,t \in T_A$ and $p \in C_A$.

Wilke [30] represented binary trees over a given alphabet A by terms over the three-sorted ranked alphabet Γ. This alphabet Γ contains operators by which A-trees and A-contexts can be constructed starting from the label alphabet A. The set of *sorts* is $S = \{\mathbf{label}, \mathbf{tree}, \mathbf{context}\}$. The *types* (see [29]) of the symbols in the S-sorted ranked alphabet $\Gamma = \{\iota, \kappa, \lambda, \rho, \eta, \sigma\}$ are as follows:

- $\iota : \mathbf{label} \to \mathbf{tree}$,
- $\lambda : \mathbf{label} \times \mathbf{tree} \to \mathbf{context}$,
- $\eta : \mathbf{context} \times \mathbf{tree} \to \mathbf{tree}$,
- $\kappa : \mathbf{label} \times \mathbf{tree} \times \mathbf{tree} \to \mathbf{tree}$,
- $\rho : \mathbf{label} \times \mathbf{tree} \to \mathbf{context}$,
- $\sigma : \mathbf{context} \times \mathbf{context} \to \mathbf{context}$.

Binary A-trees and A-contexts are represented by ΓA-terms and ΓA-contexts, as follows. If $s,t \in T_\Gamma(A)$ represent the A-trees $\hat{s}$ and $\hat{t}$, and $p,q \in C_\Gamma(A)$ represent the A-contexts $\hat{p}$ and $\hat{q}$, respectively, then for any label $a \in A$,

- $\iota(a)$ represents the A-tree c_a,
- $\lambda(a,t)$ represents the A-context $f_a(\xi, \hat{t})$,
- $\eta(p,t)$ represents the A-tree $\hat{p}(\hat{t})$, and
- $\kappa(a,s,t)$ represents the A-tree $f_a(\hat{s}, \hat{t})$,
- $\rho(a,t)$ represents the A-context $f_a(\hat{t}, \xi)$,
- $\sigma(p,q)$ represents the A-context $\hat{p}(\hat{q})$.

Any A-tree or A-context is, in general, represented by several ΓA-terms or ΓA-contexts, respectively. For example, the $\{a,b\}$-tree $f_b(f_a(c_b,c_a),c_a)$ can be represented by both of the $\Gamma\{a,b\}$-terms $\kappa(b,\kappa(a,\iota(b),\iota(a)),\iota(a))$ and $\eta(\lambda(b,\iota(a)),\kappa(a,\iota(b),\iota(a)))$.

A *Γ-algebra* $\mathcal{M} = (\langle M_\mathbf{l}, M_\mathbf{t}, M_\mathbf{c}\rangle, \Gamma)$ consists of (I) a nonempty set $M_\mathbf{l}$ of elements of sort **label**, (II) a nonempty set $M_\mathbf{t}$ of elements of sort **tree**, and (III) a nonempty set $M_\mathbf{c}$ of elements of sort **context**, and operations

- $\iota^\mathcal{M} : M_\mathbf{l} \to M_\mathbf{t}$,
- $\kappa^\mathcal{M} : M_\mathbf{l} \times M_\mathbf{t} \times M_\mathbf{t} \to M_\mathbf{t}$,
- $\lambda^\mathcal{M} : M_\mathbf{l} \times M_\mathbf{t} \to M_\mathbf{c}$,
- $\rho^\mathcal{M} : M_\mathbf{l} \times M_\mathbf{t} \to M_\mathbf{c}$,
- $\eta^\mathcal{M} : M_\mathbf{c} \times M_\mathbf{t} \to M_\mathbf{t}$, and
- $\sigma^\mathcal{M} : M_\mathbf{c} \times M_\mathbf{c} \to M_\mathbf{c}$,

defined as realizations of the symbols in Γ.

Following [30], we call a Γ-algebra a *tree algebra*, if it satisfies the following identities:

- $\sigma(\sigma(p,q),r)) \approx \sigma(p,\sigma(q,r))$;
- $\eta(\sigma(p,q),t) \approx \eta(p,\eta(q,t))$;
- $\eta(\lambda(a,s),t) \approx \kappa(a,t,s)$;
- $\eta(\rho(a,s),t) \approx \kappa(a,s,t)$.

Here, a is a variable of sort **label**, s,t are variables of sort **tree**, and p,q,r are variables of sort **context**.

The existence of a variety theorem for these tree algebras was an unanswered question (see e.g. [6, 7, 26, 30]). It was shown in Chapter 6 that such a variety theorem can not be proved for the class of tree algebras. Wilke [30] anticipated a variety theorem for **label**-generated tree algebras (i.e., the tree algebras $\mathcal{M} = (\langle M_\mathbf{l}, M_\mathbf{t}, M_\mathbf{c}\rangle, \Gamma)$ such that $\mathcal{M}$ is generated by $M_\mathbf{l}$). As a matter of fact there is no variety theorem for these tree algebras either; as the following argument shows (see [21]).

Let $\mathcal{A} = \big(\{a,b\},\{t\},\{p\},\Gamma\big)$ where $\iota^\mathcal{A}(a) = \iota^\mathcal{A}(b) = \kappa^\mathcal{A}(a,t,t) = \kappa^\mathcal{A}(b,t,t) = \eta^\mathcal{A}(p,t) = t$, and $\lambda^\mathcal{A}(a,t) = \lambda^\mathcal{A}(b,t) = \sigma^\mathcal{A}(p,p) = p$. Clearly, the algebra $\mathcal{A}$ is **label**-generated. Let $\mathbf{K}$ be the variety of tree algebras generated by $\{\mathcal{A}\}$. It can be easily shown that every member of $\mathbf{K}$ is **label**-generated, and the variety of tree languages associated with $\mathbf{K}$ has to have trivial languages only (i.e., consists of $\{\emptyset, T_A\}$ for any alphabet A). But this trivial variety must in turn be associated with the variety of trivial tree algebras; since $\mathbf{K}$ contains a non-trivial tree algebra ($\mathcal{A}$), no isomorphism can be established between varieties of finite (even **label**-generated) tree algebras and families of binary tree languages.

In Chapter 6 a variety theorem was proved for reduced tree algebras (which form a subclass of **label**-generated tree algebras) and certain families of binary tree languages [21]. I have also proved (in [17, 18]) that free tree algebras over alphabets with at least seven members are affine-complete (i.e., all congruence-preserving functions of those algebra are obtained as compositions of the constant functions, projection functions, and Wilke's functions). This is not true for free tree algebras over two-letter alphabets, and it is not known whether the theorem holds for alphabets with three to six letters.

References

[1] Almeida J., On pseudovarieties, varieties of languages, filters of congruences, pseudoidentities and related topics, *Algebra Universalis* **27** (1990), 333–350.

[2] Büchi J.R., *Finite automata, their algebras and grammar; Towards a theory of formal expressions*, Edited and with a preface by Dirk Siefkes, Springer-Verlag, New York, 1989.

[3] Comon H. et al., *Tree automata techniques and applications*, An evolving web text since 1997.
`http://www.grappa.univ-lille3.fr/tata`

[4] Denecke K. & Wismath S.L., *Universal algebra and applications in theoretical computer science*, Chapman & Hall/CRC, Boca Raton, FL, 2002.

[5] Eilenberg S., *Automata, languages, and machines*, Vol. B., Pure and Applied Mathematics, Vol. 59, Academic Press, New York, London, 1976.

[6] Ésik Z., A variety theorem for trees and theories, *Publ. Math. Debrecen* **54** (1999), 711–762.

[7] Ésik Z. & Weil P., On logically defined recognizable tree languages, in: Pandya P. K. & Radhakrishnan J. (eds.), *Proceedings of FSTTCS'03*, Lect. Notes in Comput. Sci. **2914**, Springer-Verlag (2003), 195-207.

[8] Gécseg F. & Steinby M., *Tree automata*, Akadémiai Kiadó (Publishing House of the Hungarian Academy of Sciences), Budapest, 1984.

[9] Gécseg F. & Steinby M., Tree languages, in: Rozenberg G. & Salomaa A. (eds.) *Handbook of formal languages*, Vol. 3, Springer, Berlin (1997), 1–68.

[10] Jantzen M., *Basics of term rewriting*, in: Rozenberg G. & Salomaa A. (eds.) *Handbook of formal languages*, Vol. **3**, Springer, Berlin (1997), 269–337.

[11] Nivat M. & Podelski A., Tree monoids and recognizability of sets of finite trees, in: Aït-Kaci H. & Nivat M. (eds.), *Resolution of Equations in Algebraic Structures*, Vol. 1, Academic Press, Boston MA (1989), 351–367.

[12] Nivat M. & Podelski A., Definite tree languages (cont'd), *Bull. EATCS* **38** (1989) 186–190.

[13] Petković T. & Salehi S., Positive varieties of tree languages, *Theoret. Comput. Sci.* **347** (2005), 1–35.

[14] Pin J.E., *Varieties of formal languages*, Foundations of Computer Science, North Oxford Academic Publishers, Oxford, 1986.

[15] Pin J.E., A variety theorem without complementation, Izvestiya VUZ Matematika **39** (1995), 80–90. English version, *Russian Mathem. (Iz. VUZ)* **39** (1995), 74–63.

[16] Podelski A., A monoid approach to tree languages, in: Nivat M. & Podelski A. (eds.) *Tree Automata and Languages*, Elsevier, Amsterdam (1992), 41–56.

[17] Salehi S., A completeness property of Wilke's tree algebras, in: Rovan B. & Vojtáš P. (eds.), *Proc. MFCS'2003*, Lect. Notes Comp. Sci. **2747**, Springer-Verlag (2003), 662–670.

[18] Salehi S., Congruence preserving functions of Wilke's tree algebras, *Algebra Universalis* **53** (2005), 451–470.

[19] Salehi S., Varieties of tree languages definable by syntactic monoids, *Acta Cybernetica* **17** (2005), 21–41.

[20] Salehi S. & Steinby M., Varieties of many-sorted recognizable sets, *TUCS Technical Reports* **629**, September 2004. `http://www.tucs.fi/publications/insight.php?id=tSaSt04a`

– Journal version is submitted.

[21] Salehi S. & Steinby M., Tree algebras and varieties of tree languages, *TUCS Technical Reports* **762**, March 2006. `http://www.tucs.fi/publications/insight.php?id=tSaSt06a`

– Journal version is submitted.

[22] Salomaa K., *Syntactic monoids of regular forests* (in Finnish), M.Sc. Thesis, Deptartment of Mathematics, Turku University, 1983.

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

[24] Steinby M., Syntactic algebras and varieties of recognizable sets, in: Bidoit M. & Dauchet M. (eds.), *Proc. CAAP'79* (University of Lille 1979), 226–240.

[25] Steinby M., A theory of tree language varieties, in: Nivat M. & Podelski A. (eds.) *Tree Automata and Languages*, Elsevier, Amsterdam (1992), 57–81.

[26] Steinby M., General varieties of tree languages, *Theoret. Comput. Sci.* **205** (1998), 1–43.

[27] Thérien D., Recognizable languages and congruences, *Semigroup Forum* **23** (1981), 371–373.

[28] Thomas W., Logical aspects in the study of tree languages, in: Courcelle B. (ed.), *Ninth Colloquium on Trees in Algebra and in Programming* (Proc. CAAP'84), Cambridge University Press (1984), 31–51.

[29] Wechler W., *Universal algebra for computer scientists*, EATCS Monographs on Theoretical Computer Science **25**, Springer-Verlag, Berlin, 1992.

[30] Wilke T., An algebraic characterization of frontier testable tree languages, *Theoret. Comput. Sci.* **154** (1996), 85–106.