Hierarchies of Piecewise Testable Languages

Ondřej Klíma and Libor Polák [⋆]

Department of Mathematics, Masaryk University
Janáčkovo nám 2a, 662 95 Brno, Czech Republic

Abstract. The classes of languages which are boolean combinations of languages of the form

$$A^*a_1A^*a_2A^* \dots A^*a_\ell A^*, \text{ where } a_1,\dots,a_\ell \in A,\ \ell \le k\,,$$

for a fixed $k \ge 0$, form a natural hierarchy within piecewise testable languages and have been studied in papers by Simon, Blanchet-Sadri, Volkov and others. The main issues were the existence of finite bases of identities for the corresponding pseudovarieties of monoids and generating monoids for these pseudovarieties.

Here we deal with similar questions concerning the finite unions and positive boolean combinations of the languages of the form above. In the first case the corresponding pseudovarieties are given by a single identity, in the second case there are finite bases for k equal to 1 and 2 and there is no finite basis for $k \ge 4$ (the case $k = 3$ remains open). All the pseudovarieties are generated by a single algebraic structure.

Keywords: varieties of languages, piecewise testable languages, syntactic monoid

1 Introduction

A language L over an alphabet A is called *piecewise testable* if it is a finite boolean combination of languages of the form

$$A^*a_1A^*a_2A^* \dots A^*a_\ell A^*, \text{ where } a_1,\dots,a_\ell \in A,\ \ell \ge 0\,. \quad (*)$$

A characterization of piecewise testable languages was given by Simon [18] who proved that a language L is piecewise testable if and only if its syntactic monoid is $\mathcal{J}$-trivial. Note that nowadays there exist several proofs of this deep result [1, 7, 19, 22]. See survey papers [11, 13] for more information and connections to concatenation hierarchies.

The Simon theorem was one of the first deep examples of Eilenberg's correspondence [5] between boolean varieties of languages and pseudovarieties of monoids. The correspondence uses the concept of the syntactic monoid of a language. Pin's modification [12, 13] of Eilenberg's result gives a correspondence

---

[⋆] Both authors were supported by the Ministry of Education of the Czech Republic under the project MSM 0021622409 and by the Grant no. 201/06/0936 of the Grant Agency of the Czech Republic

between positive varieties of languages and pseudovarieties of finite ordered monoids. For example, finite unions of languages of the form $(*)$ form a positive variety of languages $\mathscr{P}$ which corresponds to the pseudovariety of all ordered monoids satisfying the identity $x \leq 1$ [14].

Later on Polák [15] presented another modification of Eilenberg's correspondence: conjunctive varieties are related to pseudovarieties of finite idempotent semirings. The difference is that conjunctive varieties of languages are not closed, in general, under complements and unions and one uses a stronger notion of preimages. Note that there is a dual version. Namely, the so-called disjunctive varieties which can be obtained from conjunctive varieties in the following way: for a conjunctive variety $\mathscr{V}$, we consider the class $\mathscr{V}^{\mathsf{c}}$ of complements of languages from $\mathscr{V}$ (see [16] for more details).

Next steps were the literal varieties of languages by Ésik and Ito (see [6]), Straubing's C-varieties of languages (see [21]), and more generally Polák's D-varieties (see [16]). We make use of all mentioned modifications of Eilenberg's correspondence in this paper. We slightly modify the definition of a semiring (no zero element is postulated) and D-varieties of languages are present in our paper only implicitely.

We are interested in piecewise testable languages. If we fix a number k then we can consider boolean combinations of languages of the form $(*)$, where $\ell \leq k$. The resulting class is a boolean variety of languages, which is denoted $\mathscr{BV}_k$. These k-levels in the hierarchy of piecewise testable languages were considered by I. Simon [18], who found a simple characterization by identities for levels 1 and 2. A lot of work was done by Blanchet-Sadri [3, 4] who found a basis of identities for the level 3 and proved that there is no finite basis of identities for $k > 3$.

In this paper we consider similar levels of positive varieties $\mathscr{PV}_k$ and disjunctive varieties $\mathscr{DV}_k$ of piecewise testable languages. Here $\mathscr{PV}_k$ is formed by finite intersections of finite unions of languages of the form $(*)$, where $\ell \leq k$, and $\mathscr{DV}_k$ is formed by finite unions of languages of the form $(*)$, where $\ell \leq k$.

After introductory Sections 1 and 2 we discuss the identity problem for the pseudovarieties corresponding to $\mathscr{BV}_k$, $\mathscr{PV}_k$ and $\mathscr{DV}_k$ in Section 3. In Section 4 we obtain a characterization of $\mathscr{DV}_k$ for arbitrary $k \geq 1$, $\mathscr{PV}_1$ and $\mathscr{PV}_2$ in terms of a basis of identities for the corresponding pseudovariety of semirings (resp. monoids) and we show that such a finite basis does not exist for $\mathscr{PV}_k$ for $k > 3$. The characterization of $\mathscr{PV}_3$ by a finite basis of identities is stated as an open problem. Note that we do not reprove the results of Simon or Blanchet-Sadri.

In [23] Volkov showed that the pseudovariety of monoids corresponding to $\mathscr{BV}_k$ is generated by a single monoid for each k. He proved that in order to get this result we can use any of the three different series of monoids described by Straubing [20] and Pin [10]. In Section 5 we present an alternative proof concerning one of the series and we show that also $\mathscr{PV}_k$ and $\mathscr{DV}_k$ are generated by a single ordered monoid and by a single idempotent semiring, respectively. Finally, in the last section we state among others that the pseudovarieties $\mathbf{BV}_k$ and $\mathbf{PV}_k$ for $k \geq 4$ have no finite bases of pseudoidentities.

Due to the space limitations certain parts of proofs are placed into Appendix (see [9]).

2 Eilenberg Type Correspondences

Let A be a finite alphabet, let A^* be the set of all words over A with the operation of concatenation $\cdot$, i.e. A^* is the free monoid over A. The empty word is denoted by λ. A language over an alphabet A is a subset of A^*. A language is *regular* if it is accepted by a finite automaton – see, for instance, [2] or [13]. We will work with certain classes of (regular) languages.

2.1 Boolean and Positive Varieties of Languages

We recall here the basics concerning the Eilenberg type theorems. The boolean case was invented by Eilenberg [5] and the positive case by Pin [12].

A *boolean variety of languages* $\mathcal{V}$ associates to every finite alphabet A a class $\mathcal{V}(A)$ of regular languages over A in such a way that

- $\mathcal{V}(A)$ is closed under boolean operations (in particular $\emptyset, A^* \in \mathcal{V}(A)$),
- $\mathcal{V}(A)$ is closed under derivatives, i.e.
  $L \in \mathcal{V}(A),\ u, v \in A^*$ implies $u^{-1}Lv^{-1} = \{\, w \in A^* \mid uwv \in L \,\} \in \mathcal{V}(A)$,
- $\mathcal{V}$ is closed under preimages in morphisms, i.e.
  $f : B^* \to A^*,\ L \in \mathcal{V}(A)$ implies $f^{-1}(L) \in \mathcal{V}(B)$.

To get the notion of a *positive* variety of languages, we use in the first item only intersections and unions (not complements).

A *pseudovariety* of finite monoids is a class of finite monoids closed under submonoids, morphic images and products of finite families. Similarly for ordered monoids (see [13]).

For a regular language $L \subseteq A^*$, we define the relations $\sim_L$ and $\preceq_L$ on A^* as follows: for $u, v \in A^*$ we have

$$u \sim_L v \text{ if and only if } (\ \forall\, p, q \in A^*\ )\ (\ puq \in L \iff pvq \in L\ )\,,$$

$$u \preceq_L v \text{ if and only if } (\ \forall\, p, q \in A^*\ )\ (\ pvq \in L \implies puq \in L\ )\,.$$

The relation $\sim_L$ is a congruence on A^* of finite index (i.e. there are only finitely many classes) and the quotient structure $\mathsf{M}(L) = A^*/\sim_L$ is called the *syntactic monoid* of L.

The relation $\preceq_L$ is a preorder (i.e. a reflexive and transitive relation) on A^* and the corresponding equivalence relation is $\sim_L$. Hence $\preceq_L$ induces an order on $\mathsf{M}(L) = A^*/\sim_L$, namely: $u{\sim_L} \leq v{\sim_L}$ if and only if $u \preceq_L v$. Then we speak about a *syntactic ordered monoid* of L and we denote the structure by $\mathsf{O}(L)$.

Result 1. (Eilenberg [5], Pin [12]) *Boolean varieties (positive varieties) of languages correspond to pseudovarieties of finite monoids (ordered monoids).*

The correspondence, written $\mathscr{V} \longleftrightarrow \mathbf{V}$ *(*$\mathscr{P} \longleftrightarrow \mathbf{P}$*), is given by the following relationship: for* $L \subseteq A^*$ *we have*

$$L \in \mathscr{V}(A) \iff \mathsf{M}(L) \in \mathbf{V} \quad (\ L \in \mathscr{P}(A) \iff \mathsf{O}(L) \in \mathbf{P}\ ) .$$

The pseudovarieties can be characterized by pseudoidentities (see e.g. [2], [13]). The classes we consider here are *equational* – they are given by identities. For a set X, an *identity* is a pair $u = v$ ($u \leq v$) of words over X, i.e. $u, v \in X^*$. An identity $u = v$ ($u \leq v$) is *satisfied* in a finite monoid M (ordered monoid $(M, \leq)$) if for each morphism $\phi : X^* \to M$ we have $\phi(u) = \phi(v)$ ($\phi(u) \leq \phi(v)$). In such a case we write $M \models u = v$, and for a set of identities Π, we define $\mathsf{Mod}(\Pi) = \{\, M \mid (\ \forall\ \pi \in \Pi\ )\ M \models \pi \,\}$. For a class $\mathcal{M}$ of monoids, the meaning of $\mathcal{M} \models \Pi$ is that, for each $M \in \mathcal{M}$, we have $M \models \Pi$. Similarly for the ordered case.

2.2 Disjunctive Varieties of Languages

Conjunctive (and dually disjunctive) varieties of languages were introduced by the second author [15]. In the definition of such classes of languages union (or dually intersection) is omitted but morphisms are from a larger class. Motivated by Straubing [21] the second author [16] generalized conjunctive varieties to D-conjunctive varieties where semiring morphisms can be taken from a fixed class of morphisms D. As our application uses the non-killing morphisms from [8], we modify mentioned definitions without using a concept of D-conjunctive varieties. We even modify also the basic definition of idempotent semiring to make the presentation more clear, namely we omit a neutral element with respect to the second operation, and finally we will consider the dual version, i.e. disjunctive varieties.

An *idempotent semiring* is a structure $(S, \cdot, +, 1)$ where

- $(S, \cdot)$ is a monoid with the neutral element 1,
- $(S, +)$ is a semilattice,
- $(\ \forall\ a, b, c \in S\ )\ (\ a(b + c) = ab + ac,\ \ (a + b)c = ac + bc\ )$.

A *pseudovariety* of finite idempotent semirings is a class of finite idempotent semirings closed under sub-semirings, morphic images and finite products.

Let $A^{\cup}$ denote the set of all finite non-empty subsets of A^*. For $U, V \in A^{\cup}$, we define $U \cdot V = \{uv \mid u \in U, v \in V\}$. Then $(A^{\cup}, \cdot, \cup, \{\lambda\})$ is a free idempotent semiring over A. For sets A, B, a language $L \subseteq A^*$ and idempotent semiring morphism $f : B^{\cup} \to A^{\cup}$ we define $f^{-1}(L) = \{\, u \in B^* \mid f(u) \cap L \neq \emptyset \,\}$.

An *identity* is a pair $U = V$ of elements of $X^{\cup}$. An identity $U = V$ is *satisfied* in a finite idempotent semiring S if for each morphism $\phi : X^{\cup} \to S$ we have $\phi(U) = \phi(V)$. In such a case we write $S \models U = V$ and for a set of identities Π we define $\mathsf{Mod}(\Pi) = \{\, S \mid (\ \forall\ \pi \in \Pi\ )\ S \models \pi \,\}$.

A *disjunctive variety* of languages $\mathscr{D}$ associates to every finite alphabet A a class $\mathscr{D}(A)$ of regular languages over A in such a way that

- $A^* \in \mathscr{D}(A)$,
- $\mathscr{D}(A)$ is closed under finite unions (in particular $\emptyset \in \mathscr{D}(A)$),
- $\mathscr{D}(A)$ is closed under derivatives,
- $\mathscr{D}$ is closed under preimages in semiring morphisms, i.e. $f : B^\cup \to A^\cup,\ L \in \mathscr{D}(A)$ implies $f^{-1}(L) \in \mathscr{D}(B)$.

Let $L \subseteq A^*$ be a regular language. We define the relation $\equiv_L$ on $A^\cup$ as follows: for $U, V \in A^\cup$ we have

$$U \equiv_L V \text{ if and only if } (\ \forall\, p, q \in A^*\ )\ (\ pUq \cap L \neq \emptyset \iff pVq \cap L \neq \emptyset\ ).$$

This relation has a finite index and the quotient structure $\mathsf{S}(L) = A^\cup / \equiv_L$ is called the *syntactic semiring* of L. Notice that we are using the syntactic semiring from [15, 16] for the complement of L.

Result 2. (Polák [16]) *Disjunctive varieties of languages correspond to pseudovarieties of idempotent semirings. The correspondence, written* $\mathscr{D} \longleftrightarrow \mathbf{D}$*, is given by the following relationship: for* $L \subseteq A^*$ *we have*

$$L \in \mathscr{D}(A) \iff \mathsf{S}(L) \in \mathbf{D}\ .$$

3 Hierarchies of Piecewise Testable Languages

For a word $u = a_1 a_2 \dots a_n$, where $a_1, a_2, \dots, a_n \in A$, we define the language

$$L_u = A^* a_1 A^* \dots A^* a_n A^*$$

and we denote by $|u|$ the length of the word u, i.e. $|u| = n$. Note that the length of the empty word λ is $|\lambda| = 0$ and $L_\lambda = A^*$. For a fixed $k \geq 0$, we define the classes $\mathscr{DV}_k$, $\mathscr{PV}_k$ and $\mathscr{BV}_k$ as follows. For each finite alphabet A, we have

- $L \in \mathscr{DV}_k(A)$ if L is a finite union of languages of the form L_u, where $u \in A^*$, $|u| \leq k$;
- $L \in \mathscr{PV}_k(A)$ if L is a finite intersection of finite unions of languages of the form L_u, where $u \in A^*$, $|u| \leq k$;
- $L \in \mathscr{BV}_k(A)$ if L is a boolean combination of languages of the form L_u, where $u \in A^*$, $|u| \leq k$.

Proposition 1. *Let* $k \geq 0$. *Then*
(i) $\mathscr{DV}_k$ *is a disjunctive variety of languages,*
(ii) $\mathscr{PV}_k$ *is a positive variety of languages,*
(iii) $\mathscr{BV}_k$ *is a boolean variety of languages.*

Proof. All statements are straightforward. □

We say that a word $u = b_1 b_2 \dots b_m$, where $b_1, \dots, b_m \in A$, is a *subword* of a word $v = c_1 c_2 \dots c_n$, where $b_1, \dots, c_n \in A$, if $b_1 = c_{i_1}, \dots, b_m = c_{i_m}$ for some $1 \le i_1 < i_2 < \dots < i_m \le n$. We write $u \triangleleft v$ in this case. Note that for $u \in A^*$, we have $L_u = \{\, v \in A^* \mid u \triangleleft v \,\}$.

For $u \in A^*$, we define $\mathsf{Sub}_k(u)$ as the set of all subwords of u of length at most k and $\mathsf{Sub}(u) = \bigcup_{k \ge 0} \mathsf{Sub}_k(u)$. For $U \in A^{\cup}$, we define $\mathsf{Sub}_k(U) = \bigcup_{u \in U} \mathsf{Sub}_k(u)$.

Next, we define relations $\sim_k^A$, $\prec_k^A$ on A^* and $\equiv_k^A$ on $A^{\cup}$ as follows: for $u, v \in A^*$, $U, V \in A^{\cup}$, we have

$$
\begin{aligned}
u \sim_k^A v \quad & \text{if and only if} \quad \mathsf{Sub}_k(u) = \mathsf{Sub}_k(v)\,, \\
u \preceq_k^A v \quad & \text{if and only if} \quad \mathsf{Sub}_k(v) \subseteq \mathsf{Sub}_k(u)\,, \\
U \equiv_k^A V \quad & \text{if and only if} \quad \mathsf{Sub}_k(U) = \mathsf{Sub}_k(V)\,.
\end{aligned}
$$

We write only $\sim_k$, $\preceq_k$ and $\equiv_k$ when the alphabet A is known from the context.

The first item of the following lemma is due to Simon [17] and this is a basic step in every paper concerning piecewise testable languages (see e.g. [1, 18, 23]).

Proposition 2. *Let $k \ge 0$. Then for $L \subseteq A^*$, we have*
(i) $L \in \mathscr{B}\mathscr{V}_k$ if and only if $\sim_k \subseteq \sim_L$,
(ii) $L \in \mathscr{P}\mathscr{V}_k$ if and only if $\preceq_k \subseteq \preceq_L$,
(iii) $L \in \mathscr{D}\mathscr{V}_k$ if and only if $\equiv_k \subseteq \equiv_L$.

The proof is placed into Appendix (see [9]).

By Proposition 2 the quotient structures $A^*/\sim_k$, $A^*/\preceq_k$, $A^{\cup}/\equiv_k$ are free objects in the (equational) pseudovarieties $\mathbf{BV}_k, \mathbf{PV}_k$ and $\mathbf{DV}_k$ over the set A. Since the equivalence relations $\sim_k$ and $\equiv_k$ have finite indices the corresponding pseudovarieties are locally finite. This is not a surprise because for a given alphabet A there are only finitely many languages of the form $(*)$ with $\ell \le k$, hence $\mathscr{B}\mathscr{V}_k(A)$ ($\mathscr{P}\mathscr{V}_k(A)$ and $\mathscr{D}\mathscr{V}_k(A)$, respectively) are finite.

This proposition also solves the identity problem for the pseudovarieties $\mathbf{BV}_k$, $\mathbf{PV}_k$ and $\mathbf{DV}_k$. But the solution of the identity problem is not a solution of the membership problem. Only if we have a finite basis of identities we can test them. Our goal is to find such bases for the mentioned classes.

From another point of view the proposition solves the membership problem too. One can compute the finite free structure $A^*/\sim_k$ (or $A^*/\preceq_k$, or $A^{\cup}/\equiv_k$) and check whether the syntactic monoid (ordered monoid, semiring) of a given language L is a quotient of this free structure.

We already stated that the classes $\mathscr{B}\mathscr{V}_k$ were studied in many contributions. The first two items of the following lemma are due to Simon [18], the third can be found in [3]. The last item is proved in [4].

Result 3. *Let $k \ge 0$. Then (i) $\mathbf{BV}_1 = \mathsf{Mod}(xy = yx,\ x^2 = x)$,*
(ii) $\mathbf{BV}_2 = \mathsf{Mod}((xy)^2 = (yx)^2,\ xyzx = xyxzx)$,
(iii) $\mathbf{BV}_3 = \mathsf{Mod}((xy)^3 = (yx)^3,\ xzyxvxwy = xzxyxvxwy,\ ywxvxyz = ywxvxyxzx)$,
(iv) $\mathbf{BV}_k$ is not finitely based for $k \ge 4$.

Up to our knowledge there are no similar results for the hierarchies $\mathscr{P}\mathscr{V}_k$ and $\mathscr{D}\mathscr{V}_k$. These will be established in the next section.

4 Bases of Identities for $\mathscr{PV}_k$ and $\mathscr{DV}_k$

4.1 Disjunctive Varieties $\mathscr{DV}_k$

It is not hard to see that the disjunctive variety $\mathscr{DV}_1$ corresponds to the pseudovariety of idempotent semirings $\mathsf{Mod}(xy = x + y)$. The next theorem establishes the result for an arbitrary k. We consider

$$x_1 x_2 \dots x_{k+1} \ = \sum_{i=1}^{k+1} x_1 \dots x_{i-1} x_{i+1} \dots x_{k+1} \ . \qquad (\pi_k)$$

Theorem 1. *Let $k \geq 0$. Then $\mathbf{DV}_k = \mathsf{Mod}(\pi_k)$.*

Proof. It is easy to see that both sides of π_k have the same set of subwords of length at most k. With respect to Proposition 2 we have to show that each identity $U = V$ such that $\mathsf{Sub}_k(U) = \mathsf{Sub}_k(V)$ is a consequence of the identity π_k. If we put in the identity π_k all variables equal to 1 with exception of the variable x_1, then we obtain the identity $x_1 = x_1 + 1$. From this identity we have $xy = (x + 1)(y + 1) = xy + x + y + 1$ and more generally $u = \mathsf{Sub}(u)$ for each word u. Now using the identity π_k we can rewrite each word in $\mathsf{Sub}(u)$ by all its subwords of length at most k. Hence we obtain the identity $u = \mathsf{Sub}_k(u)$. The identity $U = \mathsf{Sub}_k(U)$ follows and from that we obtain each identity $U = V$ such that $\mathsf{Sub}_k(U) = \mathsf{Sub}_k(V)$. ☐

4.2 Positive Varieties $\mathscr{PV}_k$

We prove a certain analogue of the characterizations from Result 3.Recall that the positive variety $\mathscr{P} = \bigcup_{k \geq 0} \mathscr{PV}_k$ is characterized by the identity $x \leq 1$. When we study the positive varieties $\mathscr{PV}_k$ then it is natural to consider the classes $\mathscr{BV}_k \cap \mathscr{P}$ and ask whether the equality $\mathscr{PV}_k = \mathscr{BV}_k \cap \mathscr{P}$ holds. The inclusion $\mathscr{PV}_k \subseteq \mathscr{BV}_k \cap \mathscr{P}$ is trivial but the opposite one is much more delicate as the next result shows.

Theorem 2. *(i) $\mathscr{PV}_1 = \mathscr{BV}_1 \cap \mathscr{P}$, i.e. $\mathbf{PV}_1 = \mathsf{Mod}(xy = yx, x^2 = x,\ x \leq 1)$.*
(ii) $\mathscr{PV}_2 = \mathscr{BV}_2 \cap \mathscr{P}$, i.e. $\mathbf{PV}_2 = \mathsf{Mod}((xy)^2 = (yx)^2,\ xyzx = xyxzx,\ x \leq 1)$.
(iii) For $k \geq 3$, $\mathscr{PV}_k \neq \mathscr{BV}_k \cap \mathscr{P}$.
(iv) For $k \geq 4$, $\mathbf{PV}_k$ has no finite basis of identities.

Proof. **Part (i).** Let $u \leq v$ be an identity satisfied in $\mathbf{PV}_1$. This means that $u \preceq_1 v$, i.e. $\mathsf{Sub}_1(v) \subseteq \mathsf{Sub}_1(u)$. Then $uv \sim_1 u$ and $v \triangleleft uv$. Hence $\mathbf{BV}_1 \models u = uv$ and $\mathbf{P} \models uv \leq v$ and the identity $u \leq v$ is a consequence of identities satisfied in $\mathbf{BV}_1$ and $\mathbf{P}$. This implies $\mathscr{BV}_1 \cap \mathscr{P} \subseteq \mathscr{PV}_1$. The statement follows.

Part (ii). We start with a technical lemma which we use inductively afterwards.

Lemma 1. *Let $u, v \in A^*$ be such that $u \preceq_2 v$ and $u \not\sim_2 v$. Then there exists $w \in A^*$ such that $u \preceq_2 w \preceq_2 v$ and at least one of the following two conditions happens: i) $v \triangleleft w$ and $v \not\sim_2 w$ or ii) $w \triangleleft u$ and $w \not\sim_2 u$.*

The proof is placed into Appendix (see [9]).

Claim 1. $\mathscr{PV}_2 = \mathscr{BV}_2 \cap \mathscr{P}$.

Proof. We show that each identity $u \leq v$ which is satisfied in $\mathbf{PV}_2$ is a consequence of identities satisfied in $\mathbf{BV}_2$ and the identity $x \leq 1$. We show this by an induction with respect to the cardinality of the set $M = \mathsf{Sub}_2(u) \setminus \mathsf{Sub}_2(v)$.

If $M = \emptyset$ then $u \sim_2 v$ and the statement is clear.

If $M \neq \emptyset$ then the assumptions of Lemma 1 are valid. So, there exists w such that $u \preceq_2 w \preceq_2 v$, $v \lhd w$ and $v \not\sim_2 w$ (or $w \lhd u$ and $w \not\sim_2 u$ which can be proved in a similar way). Then $\mathsf{Sub}_2(v) \subset \mathsf{Sub}_2(w) \subseteq \mathsf{Sub}_2(u)$ and $\mathsf{Sub}_2(u) \setminus \mathsf{Sub}_2(w) \subset M$ follows. By an induction assumption the identity $u \leq w$ is a consequence of the identities satisfied in $\mathbf{BV}_2$ and the identity $x \leq 1$. Because $v \lhd w$, the identity $w \leq v$ is a consequence of the identity $x \leq 1$. This implies that $u \leq v$ is a consequence of the identities satisfied in $\mathbf{BV}_2$ and the identity $x \leq 1$. ☐

Part (iii).

Claim 2. *For $k \geq 3$, it holds $\mathscr{PV}_k \neq \mathscr{BV}_k \cap \mathscr{P}$.*

Proof. We show that the identity $(xy)^{k-1} \leq x^{k-1}y^{k-1}$ is satisfied in $\mathbf{PV}_k$ but it is not satisfied in $\mathbf{BV}_k \cap \mathbf{P}$.

The first observation is clear since $\mathsf{Sub}_k(x^{k-1}y^{k-1}) \subseteq \mathsf{Sub}_k((xy)^{k-1})$.

For the second part, we assert first that there is no word v different from $(xy)^{k-1}$ such that $v \sim_k (xy)^{k-1}$. Indeed, if $\mathsf{Sub}_k(v) = \mathsf{Sub}_k((xy)^{k-1})$ then v contains exactly $k-1$ occurrences of variable x and the same number of occurrences of y. Now $yx^{k-1} \notin \mathsf{Sub}_k(v)$ and $y^{k-1}x \notin \mathsf{Sub}_k(v)$ hence the first letter of v is x and the last letter of v is y. Moreover, $x^iyx^{k-1-i} \in \mathsf{Sub}_k(v)$ for each $i = 1, \ldots, k-2$, so, between i-th and $(i+1)$-th occurrence of x in v has to be some y. We can conclude with $v = (xy)^{k-1}$.

Now we assert that there is no proper subword v of $(xy)^{k-1}$ such that $v \preceq_k x^{k-1}y^{k-1}$. Assume that there is some word v with this property. Then from $\mathsf{Sub}_k(x^{k-1}y^{k-1}) \subseteq \mathsf{Sub}_k(v)$ we can deduce that v contains exactly $k-1$ occurrences of variable x and the same number of occurrences of y, which is a contradiction.

Our two assertions imply the statement, since there is no proof of $(xy)^{k-1} \leq x^{k-1}y^{k-1}$ using the identity $x \leq 1$ and the identities which are satisfied in $\mathbf{BV}_k$. ☐

Remark 1. The idea from our proof of Claim 2 can be also used for direct construction of a language L with the properties: $L \in \mathscr{BV}_k \cap \mathscr{P}$, $L \notin \mathscr{PV}_k$. We show such an example for the case $k = 3$.

We consider the following language L over $A = \{a, b\}$

$$\begin{aligned} L &= L_{aaa} \cup L_{bbb} \cup \{aabb\} && (1)\\ &= L_{aaa} \cup L_{bbb} \cup L_{aabb} && (2)\\ &= L_{aaa} \cup L_{bbb} \cup (L_{aa} \cap L_{bb} \cap L_{ba}^{\mathsf{c}}) \ . && (3) \end{aligned}$$

The fact $L \in \mathscr{P}$ follows from (2) and the fact $L \in \mathscr{BV}_3$ follows from (3). On the other hand, we can show that $L \notin \mathscr{PV}_3$. Assume, for a moment, that $L \in \mathscr{PV}_3$. Then $\preceq_3 \subseteq \preceq_L$ by Proposition 2. It is clear that $abab \preceq_3 aabb$, so we have $abab \preceq_L aabb$ which is a contradiction with $aabb \in L$ and $abab \notin L$.

Part (iv). This part of the theorem is proved for $k = 4$ first.

Claim 3. *There is no finite basis of identities for the pseudovariety* $\mathbf{PV}_4$*.*

Proof. Assume that $\mathbf{PV}_4$ has a finite basis Π of identities. Let n be the number of variables used in Π. We consider the identity $u \leq v$ where

$$u = x\,y\,x\,X\,y\,Y\,y\ ,\ \ v = x\,x\,y\,X\,y\,Y\,y \ \text{ with } X = z_1 z_2 \dots z_n \text{ and } Y = z_n \dots z_2 z_1\ .$$

One can show that this identity is satisfied in $\mathbf{PV}_4$ and it is not a consequence of identities in Π. A full proof is placed into appendix (see [9]). □

The previous proposition can be easily modified for every $k > 4$. The change is that we multiply the words u and v by x^{k-4} from left. Hence we have

$$u = x^{k-3}\,y\,x\,X\,y\,Y\,y\ ,\quad v = x^{k-2} y\,X\,y\,Y\,y\ .$$

This ends the sketch of the proof of Theorem 2. □

For the last case $k = 3$, the proof of Claim 3 does not work. The easiest example of the identity which is satisfied in $\mathbf{PV}_3$ but which is not a consequence of the identities from $\mathbf{BV}_3$ and the identity $x \leq 1$ is the identity $xz_1yxz_2y \leq xz_1xyz_2y$. It seems that this identity is strong enough as we did not find some identity which is not a consequence of this one. This leads us to the following conjecture about finite basis of identities for $\mathbf{PV}_3$.

Conjecture. $\mathbf{PV}_3 = \mathbf{BV}_3 \cap \mathbf{P} \cap \mathsf{Mod}(\, xz_1yxz_2y \leq xz_1xyz_2y\,)$.

5 Generating by a Single Monoid and Semiring

Volkov in [23] proved that each pseudovariety $\mathbf{BV}_k$ is generated by a single monoid. We show an alternative proof of this fact and we will prove the similar results concerning the pseudovarieties $\mathbf{PV}_k$ and $\mathbf{DV}_k$. The idea is that we will generate the varieties of languages $\mathscr{BV}_k$, $\mathscr{PV}_k$ and $\mathscr{DV}_k$ by a single language instead of generating the pseudovarieties of monoids and semirings.

Volkov used three types of monoids which were introduced by Straubing and Pin, namely the monoid $\mathcal{R}_k$ of all reflexive binary relations (viewed as a submonoid of the monoid of all $(k{+}1)\times(k{+}1)$ matrices over the Boolean semiring $\mathbf{B} = (\{0,1\}, \wedge, \vee)$), its submonoid $\mathcal{U}_k$ of all upper unitriangular matrices (i.e. there are only zeros under the main diagonal and all diagonal entries are 1), and the monoid $\mathcal{C}_k$ of all order preserving and extensive transformations of a chain with $k + 1$ elements. We identify such transformation ϕ with the matrix $C(\phi)$ having exactly one non-zero entry in each row, namely at the position $(i, \phi(i))$ for $i = 1, \dots, k + 1$. Clearly, the composition of transformations corresponds to the multiplication of matrices.

The last monoid we will use, denoted $\mathcal{S}_k$, is the submonoid of $\mathcal{U}_k$ consisting of all *stair triangular* matrices, i.e. matrices satisfying: if $a_{i,j} = 1,\ i < j$ then

$$a_{i,i} = a_{i,i+1} = \cdots = a_{i,j} = 1,\ a_{i,j} = a_{i+1,j} = \cdots = a_{j,j} = 1\ .$$

The monoids $\mathcal{R}_k, \mathcal{U}_k$ and $\mathcal{S}_k$ are idempotent semirings with respect to $\vee$ taken componentwise.

Notice that the mapping $\phi \mapsto S(\phi),\ \phi \in \mathcal{C}_k$, where $(S(\phi))_{i,j} = 1$ if and only if $j \in \{i, i+1, \ldots, \phi(i)\}$, induces a monoid isomorphism of $\mathcal{C}_k$ onto $\mathcal{S}_k$.

For each k, we fix the k-element alphabet $B = \{b_1, b_2, \ldots, b_k\}$ and the language

$$L(k,B) = B^* b_1 B^* b_2 B^* \ldots B^* b_k B^*\ .$$

A crucial property of $L(k,B)$ is the following lemma.

Lemma 2. *For every finite alphabet A and a word $u \in A^*$ of length k, there exists a morphism $f : A^* \to B^*$ such that $f^{-1}(L(k,B)) = L_u$.*

Proof. Let $u = a_1 a_2 \ldots a_k$, where $a_i \in A$. For each $a \in A$ we consider the sequence of indices $i_1 < i_2 < \cdots < i_\ell$ such that $a_{i_1} = a_{i_2} = \cdots = a_{i_\ell} = a$ and define $f(a) = b_{i_\ell} \ldots b_{i_2} b_{i_1}$.

An example for a better understanding: if $k = 8,\ A = \{c_1, \ldots, c_4\}$ and $u = c_4 c_3 c_1 c_4 c_1 c_3 c_1 c_4$ then $f : c_1 \mapsto b_7 b_5 b_3,\ c_2 \mapsto \lambda,\ c_3 \mapsto b_6 b_2,\ c_4 \mapsto b_8 b_4 b_1$.

Note that $b_i b_{i+1} \ldots b_{i+j} \lhd f(a)$ implies $j = 0$; in other terms $\mathsf{Sub}(f(a)) \cap \mathsf{Sub}(b_1 b_2 \ldots b_k) \subseteq B$. So, we have defined $f : A^* \to B^*$ morphism and we have to check that $f^{-1}(L(k,B)) = L_u$.

"$\subseteq$" : Let $w \in f^{-1}(L(k,B)),\ w = c_1 c_2 \ldots c_m$, where $c_1, \ldots c_m \in A$. Then there exist indices $j_1 < j_2 < \cdots < j_k$ such that $f(c_{j_i})$ contains b_i for all $i = 1, \ldots, k$. Hence $c_{j_i} = a_i$ for all $i = 1, \ldots, k$. This means $u = a_1 a_2 \ldots a_k \lhd w$, i.e. $w \in L_u$.

"$\supseteq$" : Now, let $w \in L_u$. Then $f(u) \lhd f(w)$. From the definition of images of letters we have $b_i \lhd f(a_i)$ for all $i = 1, \ldots, k$ and we can conclude with $b_1 b_2 \ldots b_k \lhd f(w)$, i.e. $w \in f^{-1}(L(k,B))$. □

Lemma 3. *For the language $L(k,B)$ over the alphabet B the following is true.*
(i) If a boolean variety of languages $\mathscr{B}$ satisfies $L(k,B) \in \mathscr{B}(B)$, then $\mathscr{B}\mathscr{V}_k \subseteq \mathscr{B}$.
(ii) If a positive variety of languages $\mathscr{V}$ satisfies $L(k,B) \in \mathscr{V}(B)$, then $\mathscr{P}\mathscr{V}_k \subseteq \mathscr{V}$.
(iii) If a disjunctive variety $\mathscr{D}$ satisfies $L(k,B) \in \mathscr{D}(B)$, then $\mathscr{D}\mathscr{V}_k \subseteq \mathscr{D}$.

Proof. In all cases the classes are closed under the preimages in morphisms. If we apply the previous lemma we see that for any alphabet A and the word u of length k we have $L_u \in \mathscr{B}(A)$ (and $L_u \in \mathscr{V}(A)$ and $L_u \in \mathscr{D}(A)$, respectively). The classes are also closed under derivatives since $a^{-1} L_{av} = L_v$ and $L_{va} a^{-1} = L_v$. Hence, for any alphabet A and the word u of length at most k, we have $L_u \in \mathscr{B}(A)$ (and $L_u \in \mathscr{V}(A)$ and $L_u \in \mathscr{D}(A)$, respectively). Now the statements are consequences of the definitions of the classes $\mathscr{B}\mathscr{V}_k$, $\mathscr{P}\mathscr{V}_k$, $\mathscr{D}\mathscr{V}_k$. □

Proposition 3. *For each $k \geq 1$, we have:*
(i) $\mathbf{BV}_k$ *is generated by the syntactic monoid* $B^*/\sim_{L(k,B)}$.
(ii) $\mathbf{PV}_k$ *is generated by the syntactic ordered monoid* $B^*/\preceq_{L(k,B)}$.
(iii) $\mathbf{DV}_k$ *is generated by the syntactic semiring* $B^{\cup}/\equiv_{L(k,B)}$.

Proof. It is a direct consequence of Lemma 3. □

Now we present natural models of the syntactic structures of the language $L(k,B)$. We define $\mu : B \to \mathcal{S}_k$ as follows: the only non-zero non-diagonal entry in the matrix $\mu(b_i)$ is $(\mu(b_i))_{i,i+1} = 1$ for $i = 1, \ldots, k$. This mapping naturally extends to B^* and $B^{\cup}$.

Proposition 4. *The structures* $(\mathcal{S}_k, \cdot)$, $((\mathcal{S}_k, \cdot, \leq)$ *and* $(\mathcal{S}_k, \cdot, \vee)$*, respectively) are isomorphic to the syntactic monoid (ordered syntactic monoid and syntactic semiring, respectively) of the language* $L(k,B)$.

Proof. Indeed, using the induction with respect to the lengths of words we see that the extension $\mu : B^* \to \mathcal{S}_k$ is given by $(\mu(u))_{i,j} = 1$ if and only if $i \leq j$ and $b_i \ldots b_{j-1} \lhd u$ for each $u \in A^*$.

For a matrix $S \in \mathcal{S}_k$ with non-zero entries $s_{1,1}, \ldots, s_{1,p_1}, \ldots, s_{k,k}, \ldots, s_{k,p_k}$, $s_{k+1.k+1}$, we see that $\mu(b_k \ldots b_{p_k-1} \ldots b_1 \ldots b_{p_1-1}) = S$ and thus μ is surjective.

Further, for each $u, v \in A^*$, $U, V \in A^{\cup}$, we have $u \sim_{L(k,B)} v$ if and only if $\mu(u) = \mu(v)$, $u \preceq_{L(k,B)} v$ if and only if $\mu(u) \geq \mu(v)$, and finally $U \equiv_{L(k,B)} V$ if and only if $\mu(U) = \mu(V)$. □

6 Final Remarks

Remark 2.[1] We know that the pseudovarieties $\mathbf{BV}_k$ and $\mathbf{PV}_k$, for $k \geq 4$, have no finite bases of identities. A natural question is whether there exist finite bases of pseudoidentities for these classes. (One can consult the background concerning pseudoidentities in Almeida's book [2].)

By Proposition 3 or by [23] each pseudovariety $\mathbf{BV}_k$ is generated by a single monoid and such a pseudovariety admits a finite basis of identities if and only if it admits a finite basis of pseudoidentities (see Corollary 4.3.8 in the book [2]). The same arguments can be used in the case of the pseudovarieties $\mathbf{PV}_k$. Therefore the pseudovarieties $\mathbf{BV}_k$ and $\mathbf{PV}_k$ have no finite bases of pseudoidentities.

Remark 3. Our goal was to get a better understanding of languages of level 1 in Straubing-Thérien hierarchy. We expect that some results from the present paper can be extended also to other hierarchies. For example, it could be interesting to study hierarchies based of locally testable languages, group languages or languages of the form

$$B_0^* a_1 B_1^* a_2 B_2^* \ldots a_\ell B_\ell^*, \text{ where } a_1, \ldots, a_\ell \in A,\ B_0, \ldots, B_\ell \subseteq A,\ \ell \leq k,\ k \text{ fixed}\ .$$

We also have formulated the conjecture that $\mathbf{PV}_3$ is finitely based.

---

[1] We express here our gratitude to Jorge Almeida for a discussion on the topic.

References

1. J. Almeida, Implicit operations on finite $\mathcal{J}$-trivial semigroups and a conjecture of I. Simon, *J. Pure Appl. Algebra* **69** (1990), 205–218
2. J. Almeida, *Finite Semigroups and Universal Algebra*, World Scientific, 1994
3. F. Blanchet-Sadri, Games, equations and the dot-depth hierarchy, *Comput. Math. Appl.* **18** (1989), 809–822
4. F. Blanchet-Sadri, Equations and monoids varieties of dot-depth one and two, *Theoret. Comput. Sci.* **123** (1994), 239–258
5. S. Eilenberg, *Automata, Languages and Machines,* Vol B, Academic Press, New York, 1976
6. Z. Ésik and M. Ito, Temporal logic with cyclic counting and the degree of aperiodicity of finite automata, *Acta Cybernetica* **16** (2003), 1–28,
7. P. Higgins, A proof of Simon's Theorem on piecewise testable languages, *Theoret. Comput. Sci.* **178** (1997), 257–264
8. O. Klíma and L. Polák, Classes of meet automata, *Theoret. Comput. Sci.*, to appear
9. O. Klíma and L. Polák, Hierarchies of piecewise testable languages, a version containing also Appendix, `http://www.math.muni.cz/~polak`
10. J.-E. Pin, *Varieties of Formal Languages*, North Oxford Academic, Plenum, 1986
11. J.-E. Pin, Finite semigroups and recognizable languages: an introduction, in *NATO Advanced Study Institute Semigroups, Formal Languages and Groups*, J.Fountain ed., 1–32, Kluwer Academic Publisher, 1995
12. J.-E. Pin, A variety theorem without complementation, *Russian Mathem. (Iz. VUZ)* **39** (1995) 74–83.
13. J.-E. Pin, Syntactic semigroups, Chapter 10 in *Handbook of Formal Languages*, G. Rozenberg and A. Salomaa eds, Springer, 1997
14. J.-E. Pin and P.Weil, Polynomial closure and unambiguous product, *Theory Comput. Syst.* **30** (1997), 1–39
15. L. Polák, A classification of rational languages by semilattice-ordered monoids, *Arch. Math. (Brno)* **40** (2004), 395–406
16. L. Polák, On pseudovarieties of semiring morphisms, in *Proc. MFCS 2004*, LNCS Vol. 3153 (2004), 635–647
17. I. Simon, Hierarchies of events of dot-depth one, Ph.D. thesis, University of Waterloo, 1972
18. I. Simon, Piecewise testable events, in *Proc. ICALP 1975*, LNCS Vol. 33 (1975), 214–222
19. J. Stern, Characterization of some classes of regular events, *Theoret. Comput. Sci.* **35**, (1985), 17–42
20. H. Straubing, On finite $\mathcal{J}$-trivial monoids, *Semigroup Forum* **19** (1980), 107–110
21. H. Straubing, On logical description of regular languages, in *Proc. LATIN 2002*, LNCS 2286 (2002), 528–538
22. H. Straubing and D. Thérien, Partially ordered finite monoids and a theorem of I. Simon, *J. Algebra* **119** (1988), 393–399
23. M. V. Volkov, Reflexive relations, extensive transformations and piecewise testable languages of a given height, *Internat. J. Algebra Comput.* **14** (2004), 817–827