Cantor Topologies for Finite Words *

Manfred Kufleitner Alexander Lauser

University of Stuttgart, FMI

February 24, 2011

Abstract. We consider the Cantor topology over finite words, which is obtained by resembling the well-known Cantor topology over infinite words. We consider two automata models: Complete flip automata recognize exactly the class of regular open sets, and deterministic weak automata are expressively complete for the class of regular Boolean combinations of open sets. These automata models yield decidability of the membership problem. In addition, we obtain simple and effective algebraic characterizations for regular open languages and for regular Boolean combinations of open sets. The algebraic characterizations admit counterparts for the left-right dual and the two-sided version of the Cantor topology over finite words.

As an application, we consider Boolean combinations of open sets which are recognizable by monoids in the variety **DA**. As it turns out, several characterizations of **DA**-languages admit natural restrictions for this subclass.

1 Introduction

The Cantor topology over infinite words is an important concept for classifying ω-languages. For example, every ω-regular language can be decomposed into a safety and a liveness property [1], that is, into a closed and a dense set. An ω-regular language is deterministic if and only if it is a countable intersection of open sets, cf. [19, Remark 5.1]. There are many other properties of ω-languages which can be described using the Cantor topology, see e.g. [13, 17].

For a given topology, let G be the open sets, and let F be the closed sets. Let H_σ consist of all countable unions of sets in H, and let H_δ consist of all countable intersections of sets in H. Starting with open and closed sets, alternating the operations σ and δ leads to the Borel hierarchy. From a topological point of view, an interesting property of the Cantor topology over infinite words is that it is metrizable, see [13, Proposition III.3.2]. This implies that the Boolean closure $\mathbb{B}G$ of G (or equivalently of F) in the Cantor topology over infinite words is contained in $F_\sigma \cap G_\delta$. Moreover, when restricted to ω-regular languages ω-REG, then

$$\mathbb{B}G \cap \omega\text{-REG} \;=\; F_\sigma \cap G_\delta \cap \omega\text{-REG}, \tag{1}$$

see [13, Theorem VI.3.7]. Another well-known fact is the following classification of ω-regular languages within the Borel hierarchy [19, Theorem 5.2]:

$$\omega\text{-REG} \;\subset\; \mathbb{B}(F_\sigma) \;=\; \mathbb{B}(G_\delta) \;\subseteq\; F_{\sigma\delta} \cap G_{\delta\sigma}. \tag{2}$$

*Supported by the German Research Foundation (DFG), project DI 435/5-1.

Over finite words, resembling the definition of the Cantor topology does not yield a metrizable topology. And even worse, the Borel hierarchy of the Cantor topology over finite words collapses at the very first level, i.e., $F = F_\sigma$ and $G = G_\delta$, see [17, Section 2.4]. Therefore, in view of (1) we consider the Boolean closure of Cantor sets over finite words instead of the Borel hierarchy. Our main results can be summarized as follows. For regular Cantor sets as well as for regular Boolean combinations of Cantor sets we give

- an effective and expressively complete automaton model,
- an effective algebraic characterization in terms of Green's relations,
- a single lattice equation [7], and
- closure properties.

A *flip automaton* is an automaton with no transitions from final states to non-final states. Therefore, at each final state, all minimal complete flip automata have a self-loop for every letter of the alphabet. Complete flip automata recognize exactly the class of regular Cantor sets. Weak automata have been introduced by Muller, Saoudi, and Schupp [11]. An automaton is *weak* if the states in each strongly connected component are either all final or all non-final. We show that deterministic weak automata are expressively complete for regular $\mathbb{B}G$-sets.

Another property of the Boolean closure is that for any variety of regular languages $\mathcal{V}$ we have $\mathcal{V} \cap \mathbb{B}G = \mathbb{B}(\mathcal{V} \cap G)$. Except for the automaton model, the corresponding properties for the left-right dual and the two-sided version of the Cantor topology hold as well. As for infinite words, we show that every regular language over finite words is the intersection of a closed and a dense set with respect to the Cantor topology. When building the Borel hierarchy starting with $\mathbb{B}G$, then we obtain the following counterpart of (2) for finite words: Any language is contained in $(\mathbb{B}G)_\sigma \cap (\mathbb{B}G)_\delta$.

In Section 5, we consider Boolean combinations of Cantor sets within the variety of regular languages $\mathcal{DA}$. We show that several of the combinatorial characterizations of $\mathcal{DA}$ admit natural restrictions for $\mathcal{DA} \cap \mathbb{B}G$. These characterizations include unambiguous polynomials [15], rankers [20], and partially ordered two-way automata [16]. Partially ordered automata are also known as *very weak*, *1-weak*, or *linear* automata.

2 Preliminaries

The set of finite words over the alphabet Γ is denoted by Γ^*, and the set of finite non-empty words is Γ^+. The empty word is ε. Let $L \subseteq \Gamma^*$.

- If $L\Gamma^* \subseteq L$, then L is a *Cantor set* in Γ^*.
- If $\Gamma^*L \subseteq L$, then L is a *left Cantor set* in Γ^*.
- If $\Gamma^*L\Gamma^* \subseteq L$, then L is a *two-sided Cantor set* in Γ^*.

The Cantor sets (resp. left Cantor sets, resp. two-sided Cantor sets) are the *open* sets of the *Cantor topology* (resp. the *left Cantor topology*, resp. the *two-sided Cantor topology*). A language is *closed* if its complement is open. A language $L \subseteq \Gamma^*$ is *dense* if Γ^* is the only closed set containing L.

As usual, an *automaton* $\mathcal{A} = (Q, \Gamma, \delta, Q_0, F)$ is given by a finite set of states Q, an input alphabet Γ, a transition relation $\delta \subseteq Q \times \Gamma \times Q$, a set of initial states $Q_0 \subseteq Q$, and a set of final states $F \subseteq Q$. We inductively extend the transition relation to words: $(q, \varepsilon, q) \in \delta$ for all $q \in Q$; and $(p, au, q) \in \delta$ if there exists some $r \in Q$ such that $(p, a, r) \in \delta$ and $(r, u, q) \in \delta$. We always assume that all states are reachable, i.e., for every $q \in Q$ there exist $q_0 \in Q_0$ and $u \in \Gamma^*$ such that $(q_0, u, q) \in \delta$. A word $u \in \Gamma^*$ is *accepted* by $\mathcal{A}$ if $(Q_0 \times \{u\} \times F) \cap \delta \neq \emptyset$. The language *recognized* by $\mathcal{A}$ is $L(\mathcal{A}) = \{u \in \Gamma^* \mid u \text{ is accepted by } \mathcal{A}\}$. The automaton $\mathcal{A}$ is *complete* if for

every $p \in Q$ and for every $a \in \Gamma$ there exists at least one state $q \in Q$ such that $(p, a, q) \in \delta$. And the automaton $\mathcal{A}$ is *deterministic* if $|Q_0| = 1$ and for all $p \in Q$ and all $a \in \Gamma$ there is at most one state $q \in Q$ with $(p, a, q) \in \delta$. For a deterministic automaton $\mathcal{A}$ we frequently write $\delta(p, u) = q$ instead of $(p, u, q) \in \delta$.

We view an automaton also as an edge-labeled graph with vertices Q and edges $p \xrightarrow{a} q$ for $a \in \Gamma$ given by transitions $(p, a, q) \in \delta$. A *flip automaton* is an automaton such that the intersection of $F \times \Gamma \times (Q \setminus F)$ and δ is empty, i.e., there are no transitions from final to non-final states. The idea is that in every run, flip automata can "flip" at most once from non-accepting to accepting. Note that for complete flip automata we may always assume that there is a self-loop $q \xrightarrow{a} q$ for every final state $q \in F$ and for every $a \in \Gamma$. An automaton is *weak* if for every strongly connected component $C \subseteq Q$, we either have $C \subseteq F$ or $C \cap F = \emptyset$. The concept of weak automata has been introduced by Muller, Saoudi, and Schupp [11] for alternating tree automata. Since then, weak automata have been used frequently; see e.g. [2, 4, 9, 10, 12].

For a language $L \subseteq \Gamma^*$ and a word $w \in \Gamma^*$, the left quotient of L by w is $w^{-1}L = \{u \in \Gamma^* \mid wu \in L\}$, and symmetrically, the right quotient of L by w is $Lw^{-1} = \{u \in \Gamma^* \mid uw \in L\}$. A *class of regular languages* $\mathcal{V}$ associates with every finite alphabet Γ a set $\mathcal{V}(\Gamma^*)$ of regular languages over Γ. It is closed under inverse homomorphisms if for every homomorphism $h : \Sigma^* \to \Gamma^*$, $L \in \mathcal{V}(\Gamma^*)$ implies $h^{-1}(L) \in \mathcal{V}(\Sigma^*)$. A *variety of languages* is a class of languages closed under Boolean combinations, under inverse homomorphisms, left quotients, and right quotients. In this setting, Boolean combinations consist of complementation, *finite* unions, and *finite* intersections. The closure of $\mathcal{V}$ under Boolean combinations is $\mathbb{B}\mathcal{V}$. Lattice equations can be defined in the general setting of free profinite monoids [7]. In this paper, we only introduce the ω-notation. We inductively define ω-terms over a set of variables Σ: Every $x \in \Sigma$ is an ω-term; and if x and y are ω-terms, then so are xy and $(x)^\omega$. For a number $n \in \mathbb{N}$ and an ω-term x let $x(n) \in \Sigma^*$ be the word obtained by replacing $(x)^\omega$ by the power $x^{n!}$. A regular language L satisfies the lattice equation $x \to y$ for ω-terms x and y if there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ and for all homomorphisms $h : \Sigma^* \to \Gamma^*$ the implication $h\big(x(n)\big) \in L \Rightarrow h\big(y(n)\big) \in L$ holds. It satisfies $x \leftrightarrow y$ if it satisfies both $x \to y$ and $y \to x$.

Let M be a monoid. We introduce Green's relations over M. For $x, y \in M$ let $x \leq_{\mathcal{R}} y$ (resp. $x \leq_{\mathcal{L}} y$, resp. $x \leq_{\mathcal{J}} y$) if there exist $s, t \in M$ such that $x = ys$ (resp. $x = ty$, resp. $x = tys$). Let $\leq$ be a preorder on M. A subset $I \subseteq M$ is a $\leq$-*order ideal* of M if $x \leq y \in I$ implies $x \in I$. An element $x \in M$ is *idempotent* if $x = x^2$. In every finite monoid there exists a number $\omega \geq 1$ such that x^ω is idempotent for all $x \in M$. Let $L \subseteq \Gamma^*$ be a language. A homomorphism $h : \Gamma^* \to M$ *recognizes* L if $L = h^{-1}(P)$ for some $P \subseteq M$, i.e., $u \in L$ is equivalent to $h(u) \in P$. A monoid M recognizes L if there exists a homomorphism $h : \Gamma^* \to M$ recognizing L. We define $u \equiv_L v$ if for all $s, t \in \Gamma^*$ the equivalence $sut \in L \Leftrightarrow svt \in L$ holds. The relation $\equiv_L$ is the *syntactic congruence* of L and the equivalence classes of the syntactic congruence constitute the *syntactic monoid* $\mathrm{Synt}(L)$. The *syntactic homomorphism* $h_L : \Gamma^* \to \mathrm{Synt}(L)$ is the canonical homomorphism mapping a word to its equivalence class. The syntactic monoid of a language is finite if and only if the language is regular. Moreover, every language is recognized by its syntactic homomorphism.

A *variety of monoids* $\mathbf{V}$ is a class of monoids which is closed under taking submonoids, quotients, and finitary direct products. There is a one-to-one correspondence $\mathbf{V} \leftrightarrow \mathcal{V}$ between varieties of finite monoids $\mathbf{V}$ and varieties of regular languages $\mathcal{V}$ such that $\mathcal{V}$ contains exactly those languages which are recognized by monoids in $\mathbf{V}$, see [13, Section B.2].

3 Cantor Sets

The Cantor topology over infinite words is an important concept for classifying ω-languages. One of its main properties is that the Cantor topology over infinite words is metrizable. This is not true for the Cantor topology over finite words, see e.g. [17, Section 2.4]. On the other hand, many interesting properties over finite words can be stated as follows: There exists a prefix which has some desirable property L and we do not care about subsequent actions. This immediately leads to the Cantor set $L\Gamma^*$. Below, we give simple characterizations of regular Cantor sets in terms of automata and homomorphisms onto finite monoids.

Theorem 1. *Let $L \subseteq \Gamma^*$ be a regular language. The following are equivalent:*

1. *L is a Cantor set.*
2. *L satisfies the lattice equation $y \to yz$.*
3. *Every complete deterministic automaton recognizing L is a flip automaton.*
4. *There exists a complete (nondeterministic) flip automaton recognizing L.*
5. *$h(L)$ is a $\leq_{\mathcal{R}}$-order ideal for every surjective homomorphism $h : \Gamma^* \to M$ recognizing L.*
6. *$h_L(L)$ is a $\leq_{\mathcal{R}}$-order ideal for the syntactic homomorphism $h_L : \Gamma^* \to \mathrm{Synt}(L)$.*
7. *There exists a homomorphism $h : \Gamma^* \to M$ recognizing L such that $L = h^{-1}(P)$ for some $\leq_{\mathcal{R}}$-order ideal P.*

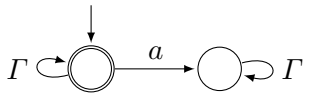
Proof. "1 $\Rightarrow$ 2": Let $x \in L$ and $y \in \Gamma^*$. We have $xy \in L\Gamma^* \subseteq L$ and therefore, L satisfies the lattice equation $y \to yz$.

"2 $\Rightarrow$ 3": Let $\mathcal{A} = (Q, \Gamma, \delta, q_0, F)$ be a complete deterministic automaton recognizing L and suppose $\delta(q_0, u) = q \in F$. Since $\mathcal{A}$ is complete, for every $v \in \Gamma^*$ there exists $p \in Q$ such that $\delta(q, v) = p$. Moreover $p \in F$, because $uv \in L$. Therefore, $\mathcal{A}$ is a flip automaton. The implication "3 $\Rightarrow$ 4" is trivial.

"4 $\Rightarrow$ 5": Let $\mathcal{A} = (Q, \Gamma, \delta, Q_0, F)$ be a complete flip automaton recognizing L and let $h : \Gamma^* \to M$ be a surjective homomorphism recognizing L. Suppose $h(w) \leq_{\mathcal{R}} h(u)$ and $h(u) \in h(L)$. Then there exists $v \in \Gamma^*$ such that $h(w) = h(uv)$. We have $(q_0, u, q) \in \delta$ for some initial state q_0 and some final state q. Since $\mathcal{A}$ is complete, there exists a state p with $(q, v, p) \in \delta$. And since $\mathcal{A}$ is a flip automaton, the state p is a final state. Therefore $uv \in L(\mathcal{A}) = L$ and $h(w) = h(uv) \in h(L)$. The implications "5 $\Rightarrow$ 6" and "6 $\Rightarrow$ 7" are trivial.

"7 $\Rightarrow$ 1": We have $h(uv) \leq_{\mathcal{R}} h(u)$ for all $u, v \in \Gamma^*$. Therefore, if $u \in L$, then $uv \in L$, i.e., $L\Gamma^* \subseteq L$. □

Example 1. Not every complete nondeterministic automaton, which recognizes a Cantor set has to be a flip automaton. The complete automaton below recognizes Γ^*, but it is not a flip automaton.



In particular, condition "3" in Theorem 1 does not hold for complete nondeterministic automata. □

The next theorem is the left-right dual of Theorem 1. For the automaton characterization we would have to consider automata which read the input from right to left. We therefore omit this characterization.

Theorem 2. *Let $L \subseteq \Gamma^*$ be a regular language. The following are equivalent:*

1. *L is a left Cantor set.*
2. *L satisfies the lattice equation $y \to xy$.*
3. *$h(L)$ is a $\leq_{\mathcal{L}}$-order ideal for every surjective homomorphism $h : \Gamma^* \to M$ recognizing L.*
4. *$h_L(L)$ is a $\leq_{\mathcal{L}}$-order ideal for the syntactic homomorphism $h_L : \Gamma^* \to \mathrm{Synt}(L)$.*
5. *There exists a homomorphism $h : \Gamma^* \to M$ recognizing L such that $L = h^{-1}(P)$ for some $\leq_{\mathcal{L}}$-order ideal P.* □

For the two-sided version of the Cantor topology over finite words, we have the following characterizations.

Theorem 3. *Let $L \subseteq \Gamma^*$ be a regular language. The following are equivalent:*

1. *L is a two-sided Cantor set.*
2. *L satisfies the lattice equation $y \to xyz$.*
3. *$h(L)$ is a $\leq_{\mathcal{J}}$-order ideal for every surjective homomorphism $h : \Gamma^* \to M$ recognizing L.*
4. *$h_L(L)$ is a $\leq_{\mathcal{J}}$-order ideal for the syntactic homomorphism $h_L : \Gamma^* \to \mathrm{Synt}(L)$.*
5. *There exists a homomorphism $h : \Gamma^* \to M$ recognizing L such that $L = h^{-1}(P)$ for some $\leq_{\mathcal{J}}$-order ideal P.*

Proof. The proof is similar to the proof of Theorem 1. "1 ⇒ 2": If $y \in L$, then $xyz \in \Gamma^* L \Gamma^* \subseteq L$. Therefore, L satisfies the lattice equation $y \to xyz$.

"2 ⇒ 3": Let $h : \Gamma^* \to M$ be a surjective homomorphism recognizing L and let $h(v) \in h(L)$. For $u, w \in \Gamma^*$, the lattice equation yields $uvw \in L$. Therefore $h(uvw) \in h(L)$, showing that $h(L)$ is a $\leq_{\mathcal{J}}$-order ideal.

The implications "3 ⇒ 4" and "4 ⇒ 5" are trivial.

"5 ⇒ 1": For all $u, v, w \in \Gamma^*$ we have $h(uvw) \leq_{\mathcal{J}} h(v)$. Therefore, $v \in L$ implies uvw in L, i.e., $\Gamma^* L \Gamma^* \subseteq L$. □

These characterizations of regular open sets immediately give the following decidability result.

Corollary 1. *It is decidable whether a regular language $L \subseteq \Gamma^*$ is a Cantor set (resp. a left Cantor set, resp. a two-sided Cantor set).*

Proof. Let L be given by a complete deterministic automaton $\mathcal{A}$. For deciding whether L is a Cantor set, it suffices to check if $\mathcal{A}$ is a flip automaton by conditions "3" and "4" of Theorem 1. In order to decide whether L is a left Cantor set, we first compute a complete deterministic automaton $\mathcal{B}$ for the reversal of L, and then check whether $\mathcal{B}$ is a flip automaton. A language is a two-sided Cantor set if and only if it is both a Cantor set and a left Cantor set. Therefore, decidability for two-sided Cantor sets follows by corresponding results for Cantor sets and left Cantor sets.

Another approach is to compute the syntactic homomorphism of L and then verify whether condition "6" in Theorem 1 (resp. condition "4" in Theorem 2, resp. condition "4" in Theorem 3) holds. □

Corollary 2. *The class of regular Cantor sets (resp. regular left Cantor sets, resp. regular two-sided Cantor sets) is closed under finite unions, finite intersections, inverse homomorphisms, and left quotients (resp. right quotients for left Cantor sets, resp. no quotients for two-sided Cantor sets).*

Proof. Let $\mathcal{W}$ be the class of all regular Cantor sets (resp. left Cantor sets, resp. two-sided Cantor sets). Closure under finite union, finite intersection and inverse homomorphisms is an immediate consequence of the fact that we have a description of $\mathcal{W}$ in terms of lattice equations [7]. Let now $L \in \mathcal{W}(\Gamma^*)$. By Theorem 1, L satisfies the lattice equation $x \to xy$. Suppose $u \in a^{-1}L$, i.e., $au \in L$. The lattice equation yields $auv \in L$, and hence, $uv \in a^{-1}L$. Therefore, $a^{-1}L$ satisfies the lattice equation $x \to xy$. Hence, $a^{-1}L \in \mathcal{W}(\Gamma^*)$ by Theorem 1. Closure under right quotients for left Cantor sets is symmetric. □

Example 2. The class of Cantor sets is not closed under right quotients. For example, $L = ab\{a,b\}^*$ is a Cantor set in $\{a,b\}^*$, whereas $L\,b^{-1} = \{a\} \cup L$ is not a Cantor set. ◊

Over infinite words every ω-regular language is the intersection of a safety and a liveness property [1]. Safety properties are the closed sets, and liveness properties are dense sets. In the following theorem, we give the analogous result for finite words.

Theorem 4. *Every regular language $L \subseteq \Gamma^*$ is an intersection of regular sets $C, D \subseteq \Gamma^*$ such that C is closed and D is dense in the Cantor topology (resp. left Cantor topology, resp. two-sided Cantor topology).*

Proof. We first consider the Cantor topology. Let $h : \Gamma^* \to M$ be a surjective homomorphism recognizing L, and let $P = h(L)$. We define

$$Q = \{x \in M \mid y \leq_{\mathcal{R}} x \text{ for some } y \in P\}$$

and $R = P \cup R'$, with R' containing exactly one element from each $\leq_{\mathcal{R}}$-minimal $\mathcal{R}$-class such that if $x \,\mathcal{R}\, y \in P$ and $x \in R'$, then $x \in P$, i.e., if P contains elements from some minimal $\mathcal{R}$-class, then we choose one of those as representatives in R'. Now, $P = Q \cap R$.

The set $M \setminus Q$ is a $\leq_{\mathcal{R}}$-order ideal. By Theorem 1, $C = h^{-1}(Q)$ is closed in the Cantor topology. We claim that $D' = h^{-1}(R')$ is dense. Then $D = h^{-1}(R)$ is dense, too. To prove the claim, consider some closed set $T \subseteq \Gamma^*$ with $D' \subseteq T$. Assume that $u \in S = \Gamma^* \setminus T$. Choose $x \in R'$ such that $x \leq_{\mathcal{R}} h(u)$. Then there exists $v \in \Gamma^*$ such that $x = h(uv)$. Now $uv \in S \cap T$, since S is open and $h^{-1}(x) \subseteq D' \subseteq T$. This is a contradiction. Hence, $S = \emptyset$ and $T = \Gamma^*$. Therefore, D' and D are dense. By construction, we have $L = C \cap D$.

The proof for the left Cantor topology is left-right dual. For the two-sided Cantor topology we choose $Q = \{x \in M \mid y \leq_{\mathcal{J}} x \text{ for some } y \in P\}$ and $R = P \cup R'$, with R' containing exactly one element from each $\leq_{\mathcal{J}}$-minimal $\mathcal{J}$-class such that if P contains elements from some minimal $\mathcal{J}$-class, then we choose one of those as representatives in R'. As before, we have $P = Q \cap R$. Showing that $h^{-1}(Q)$ is closed in the two-sided Cantor topology and showing that $h^{-1}(R)$ is dense in the two-sided Cantor topology follows the lines as for the Cantor topology. □

The following proposition shows that the Borel hierarchy over Boolean combinations of open sets in Γ^* collapses at the second level, i.e., every language is contained in $(\mathbb{B}G)_\sigma \cap (\mathbb{B}G)_\delta$.

Proposition 1. *Every language is both a countable union and a countable intersection of Boolean combinations of Cantor sets (resp. left Cantor sets, resp. two-sided Cantor sets).*

Proof. It suffices to prove the claim for two-sided Cantor sets. Let $L \subseteq \Gamma^*$. We have

$$\{u\} = \Gamma^* u \Gamma^* \setminus (\Gamma^* u \Gamma^+ \cup \Gamma^+ u \Gamma^*),$$

and $L = \bigcup_{u \in L} \{u\}$ as well as $L = \bigcap_{u \notin L} \Gamma^* \setminus \{u\}$. Therefore, $L \in (\mathbb{B}G)_\sigma \cap (\mathbb{B}G)_\delta$ over the two-sided Cantor topology G. □

4 The Boolean Closure of Cantor Topologies

The Borel hierarchy over the Cantor topology for infinite words leads to a well-known classification of ω-languages. When restricted to ω-regular languages, the Boolean closure of the Cantor topology coincides with the level $F_\sigma \cap G_\delta$ (also called Δ_2) of the Borel hierarchy, see [13, Theorem VI.3.7]. For finite words, the Borel hierarchy over the Cantor topology collapses at the very first level, and the Boolean closure of the Cantor topology contains strictly more languages. In this section, we show that regular languages in the Boolean closure $\mathbb{B}G$ of the Cantor topology G admit simple and effective automata-theoretic and algebraic characterizations. Moreover, for any variety of regular languages $\mathcal{V}$ the language classes $\mathcal{V} \cap \mathbb{B}G$ and $\mathbb{B}(\mathcal{V} \cap G)$ coincide.

Theorem 5. *Let $L \subseteq \Gamma^*$ be a regular language. The following are equivalent:*

1. *L is a Boolean combination of Cantor sets in Γ^*.*
2. *L satisfies the lattice equation $z(xy)^\omega \leftrightarrow z(xy)^\omega x$.*
3. *L satisfies the lattice equation $z(xy)^\omega \to z(xy)^\omega x$.*
4. *Every deterministic automaton recognizing L is weak.*
5. *L is recognized by some deterministic weak automaton.*
6. *$h(L)$ is a union of $\mathcal{R}$-classes for every surjective homomorphism $h : \Gamma^* \to M$ recognizing L.*
7. *$h_L(L)$ is a union of $\mathcal{R}$-classes for the syntactic homomorphism $h_L : \Gamma^* \to \mathrm{Synt}(L)$.*
8. *There exists a homomorphism $h : \Gamma^* \to M$ recognizing L such that $L = h^{-1}(P)$ for some union of $\mathcal{R}$-classes P.*

Proof. We show "$1 \Rightarrow 6 \Rightarrow 7 \Rightarrow 8 \Rightarrow 1$" and "$7 \Rightarrow 2 \Rightarrow 3 \Rightarrow 4 \Rightarrow 5 \Rightarrow 6$".

"$1 \Rightarrow 6$": Let $L = \bigcup_{i=1}^{n} P_i \setminus Q_i$ with $P_i\Gamma^* \subseteq P_i$ and $Q_i\Gamma^* \subseteq Q_i$. Consider u, v such that $h(u)\ \mathcal{R}\ h(v)$, i.e., there exist $s, t \in \Gamma^*$ with $h(v) = h(us)$ and $h(u) = h(vt)$. Suppose $u \in L$. Let $u_j = u(st)^j$ and $v_j = u_j s$. Now, $h(u_j) = h(u)$, $h(v_j) = h(v)$, and u_j is a prefix of v_j which in turn is a prefix of u_{j+1}. For every u_j there exists $i \in \{1, \ldots, n\}$ such that $u_j \in P_i \setminus Q_i$. Hence by the pigeonhole principle, there exist $j < k$ with $u_j, u_k \in P_i \setminus Q_i$ for some $i \in \{1, \ldots, n\}$. Then $v_j \in P_i\Gamma^* \subseteq P_i$. If $v_j \in Q_i$, then $u_k \in Q_i\Gamma^* \subseteq Q_i$ and $u_k \notin P_i \setminus Q_i$. We conclude $v_j \notin Q_i$ and $v_j \in P_i \setminus Q_i \subseteq L$. Hence, $h(v) = h(v_j) \in h(L)$. This shows that $h(L)$ is a union of $\mathcal{R}$-classes. The implications "$6 \Rightarrow 7$" and "$7 \Rightarrow 8$" are trivial.

"$8 \Rightarrow 1$": By Theorem 1, the inverse image $h^{-1}(Q)$ of every $\leq_{\mathcal{R}}$-order ideal Q is a Cantor open set. Moreover, the inverse image of every $\mathcal{R}$-class is a Boolean combination of such languages $h^{-1}(Q)$.

"$7 \Rightarrow 2$": Let $h : \Gamma^* \to M$ be a surjective homomorphism recognizing L such that $h(L)$ is a union of $\mathcal{R}$-classes. Choose $n_0 \geq 1$ such that s^{n_0} is idempotent for every $s \in M$. Since $h\big(z(xy)^{n!}\big)\ \mathcal{R}\ h\big(z(xy)^{n!}x\big)$ for every $x, y, z \in \Gamma^*$ and for every $n \geq n_0$, we have $z(xy)^{n!} \in L$ if and only if $z(xy)^{n!}x \in L$. Therefore, L satisfies the lattice equation $z(xy)^\omega \leftrightarrow z(xy)^\omega x$. The implication "$2 \Rightarrow 3$" is trivial.

"$3 \Rightarrow 4$": Let $\mathcal{A} = (Q, \Gamma, \delta, q_0, F)$ be a deterministic automaton recognizing L. Assume $\delta(q, x) = p$ and $\delta(p, y) = q$ for $q \in F$ and $p \notin F$. Choose $z \in \Gamma^*$ such that $\delta(q_0, z) = q$. Then for all $n \in \mathbb{N}$ we have $z(xy)^n \in L$ and $z(xy)^n x \notin L$. This contradicts the lattice equation. Hence, $\mathcal{A}$ is weak. The implication "$4 \Rightarrow 5$" is trivial.

"$5 \Rightarrow 6$": Let L be recognized by some deterministic weak automaton $\mathcal{A} = (Q, \Gamma, \delta, q_0, F)$ and let $h : \Gamma^* \to M$ be a surjective homomorphism recognizing L. Choose $n \in \mathbb{N}$ such that, for all $u \in \Gamma^*$ and for all $q \in Q$, the states $\delta(q, u^n)$ and $\delta(q, u^{n+1})$ are always in the same strongly connected component of $\mathcal{A}$. Suppose $h(u)\ \mathcal{R}\ h(v)$, i.e., $h(ux) = h(v)$ and $h(vy) = h(u)$ for some $x, y \in \Gamma^*$. By construction, $\delta\big(q_0, u(xy)^n\big)$ and $\delta\big(q_0, u(xy)^{n+1}\big)$ are in the same connected component of $\mathcal{A}$.

Hence, $\delta\big(q_0, u(xy)^n\big)$ and $\delta\big(q_0, u(xy)^n x\big)$ are also in the same connected component. Therefore, we have

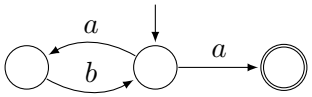
$$u \in L \text{ iff } u(xy)^n \in L \text{ iff } u(xy)^n x \in L \text{ iff } v \in L.$$

Here, the first and the third equivalence hold, since h recognizes L, and the second equivalence holds because $\mathcal{A}$ is weak. This shows that $h(L)$ is a union of $\mathcal{R}$-classes. □

Example 3. A monoid is *$\mathcal{R}$-trivial* if $\mathcal{R}$ is the identity relation. A language recognized by a finite $\mathcal{R}$-trivial monoid is a Boolean combination of Cantor sets. The converse is not true. For example, $L = \{a,b\}^* b(aa)^* b \{a,b\}^*$ is a Cantor set in $\{a,b\}^*$ (and a Boolean combination of Cantor sets over any alphabet Γ with $a, b \in \Gamma$), but L cannot be recognized by any finite aperiodic monoid. Remember that a finite monoid is aperiodic if $\mathcal{R} \cap \mathcal{L}$ is the identity relation. In particular, all finite $\mathcal{R}$-trivial monoids are aperiodic. ◊

Example 4. The language $L = (ab)^*a$ cannot be written as a Boolean combination of Cantor sets $\mathbb{B}G$. We have $(ab)^\omega a \,\mathcal{R}\, (ab)^\omega$ for every finite monoid M and for all $a, b \in M$. On the other hand, for every $n \in \mathbb{N}$ we have $(ab)^n a \in L$ and $(ab)^n \notin L$. Therefore, if $h : \Gamma^* \to M$ recognizes L, then $h(L)$ is not a union of $\mathcal{R}$-classes. By Theorem 5, it follows that L is not in $\mathbb{B}G$. ◊

Remark 1. We cannot use nondeterministic weak automata in condition "5" of Theorem 5. The automaton below is weak but it recognizes the language $(ab)^*a$ from Example 4 which is not a Boolean combination of Cantor sets.



□

In the following theorem, we give the left-right dual of Theorem 5. The automata-theoretic characterizations are omitted, since they would result in deterministic weak right-to-left automata.

Theorem 6. *Let $L \subseteq \Gamma^*$ be a regular language. The following are equivalent:*

1. *L is a Boolean combination of left Cantor sets in Γ^*.*
2. *L satisfies the lattice equation $(ts)^\omega z \leftrightarrow s(ts)^\omega z$.*
3. *L satisfies the lattice equation $(ts)^\omega z \to s(ts)^\omega z$.*
4. *$h(L)$ is a union of $\mathcal{L}$-classes for every surjective homomorphism $h : \Gamma^* \to M$ recognizing L.*
5. *$h_L(L)$ is a union of $\mathcal{L}$-classes for the syntactic homomorphism $h_L : \Gamma^* \to \mathrm{Synt}(L)$.*
6. *There exists a homomorphism $h : \Gamma^* \to M$ recognizing L such that $L = h^{-1}(P)$ for some union of $\mathcal{L}$-classes P.* □

A similar proof as for Theorem 5 leads to the following characterization of regular Boolean combinations of two-sided Cantor sets.

Theorem 7. *Let $L \subseteq \Gamma^*$ be a regular language. The following are equivalent:*

1. *L is a Boolean combination of two-sided Cantor sets in Γ^*.*
2. *L satisfies the lattice equation $(ts)^\omega z (xy)^\omega \leftrightarrow s(ts)^\omega z (xy)^\omega x$.*
3. *L satisfies the lattice equation $(ts)^\omega z (xy)^\omega \to s(ts)^\omega z (xy)^\omega x$.*
4. *$h(L)$ is a union of $\mathcal{J}$-classes for every surjective homomorphism $h : \Gamma^* \to M$ recognizing L.*
5. *$h_L(L)$ is a union of $\mathcal{J}$-classes for the syntactic homomorphism $h_L : \Gamma^* \to \mathrm{Synt}(L)$.*
6. *There exists a homomorphism $h : \Gamma^* \to M$ recognizing L such that $L = h^{-1}(P)$ for some union of $\mathcal{J}$-classes P.*

Proof. We show "$1 \Rightarrow 4 \Rightarrow 5 \Rightarrow 6 \Rightarrow 1$" and "$6 \Rightarrow 2 \Rightarrow 3 \Rightarrow 4$".

"$1 \Rightarrow 4$": Let $L = \bigcup_{i=1}^{n} P_i \setminus Q_i$ with $\Gamma^* P_i \Gamma^* \subseteq P_i$ and $\Gamma^* Q_i \Gamma^* \subseteq Q_i$. Consider u, v such that $h(u)\ \mathcal{J}\ h(v)$, i.e., there exist $s, t, r, w \in \Gamma^*$ with $h(v) = h(sut)$ and $h(u) = h(rvw)$. Assume $u \in L$. Let $u_j = (rs)^j u (tw)^j$ and $v_j = s u_j t$. Now, $h(u_j) = h(u)$, $h(v_j) = h(v)$, and u_j is a factor of v_j which in turn is a factor of u_{j+1}. For every u_j there exists $i \in \{1, \ldots, n\}$ such that $u_j \in P_i \setminus Q_i$. Hence by the pigeonhole principle, there exist $j < k$ with $u_j, u_k \in P_i \setminus Q_i$ for some $i \in \{1, \ldots, n\}$. Then $v_j \in \Gamma^* P_i \Gamma^* \subseteq P_i$. If $v_j \in Q_i$, then $u_k \in \Gamma^* Q_i \Gamma^* \subseteq Q_i$ and $u_k \notin P_i \setminus Q_i$. We conclude $v_j \notin Q_i$ and $v_j \in P_i \setminus Q_i \subseteq L$. Hence, $h(v) = h(v_j) \in h(L)$. This shows that $h(L)$ is a union of $\mathcal{J}$-classes.

The implications "$4 \Rightarrow 5$" and "$5 \Rightarrow 6$" are trivial.

"$6 \Rightarrow 1$": By Theorem 3, the inverse image $h^{-1}(Q)$ of every $\leq_{\mathcal{J}}$-order ideal Q is a two-sided Cantor set. Moreover, the inverse image of every $\mathcal{J}$-class is a Boolean combination of such languages.

"$6 \Rightarrow 2$": Let $h : \Gamma^* \to M$ be surjective recognizing L such that $h(L)$ is a union of $\mathcal{J}$-classes. Choose $n_0 \geq 1$ such that s^{n_0} is idempotent for all $s \in M$ and let $n \geq n_0$. We have $h\big((ts)^n z (xy)^n\big)\ \mathcal{J}\ h\big(s(ts)^n z (xy)^n x\big)$ for all $s, t, x, y, z \in \Gamma^*$. Thus $(ts)^n z (xy)^n \in L$ if and only if $s(ts)^n z (xy)^n x \in L$, showing the lattice equation. The implication "$2 \Rightarrow 3$" is trivial.

"$3 \Rightarrow 4$": Let $h : \Gamma^* \to M$ be a surjective homomorphism recognizing L. Suppose $h(w)\ \mathcal{J}\ h(z) \in h(L)$. Then there exist $s, t, x, y \in \Gamma^*$ such that $h(z) = h(twy)$ and $h(w) = h(szx)$. We have $h(z) = h\big((ts)^n z (xy)^n\big)$ for all $n \in \mathbb{N}$. Hence $(ts)^n z (xy)^n \in L$, because h recognizes L. Choosing n sufficiently large, the lattice equation yields $s(ts)^n z (xy)^n x \in L$. Moreover, $h(w) = h\big(s(ts)^n z (xy)^n x\big)$ and thus, $h(w) \in h(L)$. This shows that $h(L)$ is a union of $\mathcal{J}$-classes. □

Remark 2. For finite monoids, $\mathcal{J}$ is the smallest equivalence relation such that $\mathcal{R} \subseteq \mathcal{J}$ and $\mathcal{L} \subseteq \mathcal{J}$, see e.g. [13, Proposition A.2.5 (2)]. Therefore, we conclude from Theorems 5, 6, and 7 that a regular language is a Boolean combination of two-sided Cantor sets in Γ^* if and only if it is both a Boolean combination of Cantor sets in Γ^* and a Boolean combination of left Cantor sets in Γ^*. □

Corollary 3. *It is decidable whether a regular language $L \subseteq \Gamma^*$ is a Boolean combination of Cantor sets (resp. left Cantor sets, resp. two-sided Cantor sets).*

Proof. Let L be given by a deterministic automaton $\mathcal{A}$. We can effectively verify whether $\mathcal{A}$ is weak. By conditions "4" and "5" in Theorem 5, this is equivalent to L being a Boolean combination of Cantor sets. In order to decide whether L is a left Cantor set, we first compute a deterministic automaton $\mathcal{B}$ for the reversal of L and then check whether $\mathcal{B}$ is weak. For deciding whether L is a Boolean combination of two-sided Cantor sets, we verify that both conditions hold by Remark 2.

Alternatively, we can compute the syntactic homomorphism of L and then verify whether condition "7" in Theorem 5 (resp. condition "5" in Theorem 6, resp. condition "5" in Theorem 7) holds. □

Corollary 4. *The class of regular Boolean combinations of Cantor sets (resp. left Cantor sets, resp. two-sided Cantor sets) is closed under Boolean operations, inverse homomorphisms, and left quotients (resp. right quotients for left Cantor sets, resp. no quotients for two-sided Cantor sets).*

Proof. The closure properties follow from Corollary 2, since inverse homomorphisms and quotients commute with Boolean operations. □

Example 5. Let $\Gamma = \{a, b\}$. The language $L = (ab)^*aa$ is a Boolean combination of Cantor sets, because $L = L\Gamma^* \setminus L\Gamma^+$. The quotient $La^{-1} = (ab)^*a$ is the language from Example 4 which cannot be written as a Boolean combination of Cantor sets. Therefore, in general, the class of Boolean combinations of Cantor sets is not closed under right quotients. $\Diamond$

In any regular Boolean combination of Cantor sets, we can assume that the Cantor sets themselves are regular. The following theorem shows that this holds relative to any variety of regular languages.

Theorem 8. *Let $\mathcal{V}$ be a variety of regular languages and let $L \subseteq \Gamma^*$. The following assertions are equivalent:*

1. *L is a Boolean combination of Cantor sets (resp. left Cantor sets, resp. two-sided Cantor sets) in $\mathcal{V}(\Gamma^*)$.*
2. *$L \in \mathcal{V}(\Gamma^*)$ and L is a Boolean combination of Cantor sets (resp. left Cantor sets, resp. two-sided Cantor sets) in Γ^*.*

Proof. The implication "1 $\Rightarrow$ 2" is trivial since $\mathcal{V}$ is closed under Boolean operations. For "2 $\Rightarrow$ 1" let L be a Boolean combination of Cantor sets. By Theorem 5 there is a surjective homomorphism $h : \Gamma^* \to M \in \mathbf{V}$ recognizing L such that $h(L)$ is a union of $\mathcal{R}$-classes. Theorem 1 shows that the inverse image $h^{-1}(Q)$ of every $\leq_{\mathcal{R}}$-order ideal Q is a Cantor set in $\mathcal{V}(\Gamma^*)$. Since the inverse image of every $\mathcal{R}$-class is a Boolean combination of such languages $h^{-1}(Q)$, we see that L is a Boolean combination of Cantor sets in $\mathcal{V}(\Gamma^*)$. The case of being a Boolean combination of left Cantor sets follows by left-right symmetry and the proof for two-sided Cantor sets is similar. $\square$

As we will see in Section 5, what essentially happens in Boolean combinations of Cantor sets is that the end of words is "concealed." Therefore as shown in the following proposition, one way of obtaining Boolean combinations of Cantor sets is to append a new unique last symbol c. This way, the end of words is not "concealed" anymore. In particular, if $\mathcal{V}$ is a variety of regular languages and if $Lc \in \mathcal{V}(\Gamma^*)$ for $L \subseteq (\Gamma \setminus \{c\})^*$, then by Theorem 8 the language Lc is a Boolean combination of Cantor sets in $\mathcal{V}(\Gamma^*)$.

Proposition 2. *Let $L \subseteq (\Gamma \setminus \{c\})^*$ and $c \in \Gamma$. Then Lc is a Boolean combination of Cantor sets in Γ^*.*

Proof. Then $Lc = Lc\Gamma^* \setminus Lc\Gamma^+$. Hence, Lc is a Boolean combination of Cantor sets in Γ^*. $\square$

5 Boolean Combinations of Cantor Sets in $\mathcal{DA}$

A finite monoid M belongs to the variety **DA** if $(xy)^\omega x(xy)^\omega = (xy)^\omega$ for all $x, y \in M$. The corresponding variety of regular languages is denoted by $\mathcal{DA}$. It has a huge number of equivalent characterizations, see e.g. [18, 5]. In this section, we consider Boolean combinations of Cantor sets in $\mathcal{DA}$. As it turns out, several descriptions of languages in $\mathcal{DA}$ admit natural restrictions characterizing Boolean combinations of Cantor sets in $\mathcal{DA}$. Among these characterizations are unambiguous polynomials [15], partially ordered two-way automata [16], and rankers [20].

A *one-pass two-way automaton* $\mathcal{A} = (Q, \Gamma, \delta, X_0, F)$ is given by a finite set of states $Q = X \dot\cup Y$, an input alphabet Γ, a transition relation $\delta \subseteq (Q \times \Gamma \times Q) \cup (Y \times \{\triangleright\} \times X)$, a set of initial states $X_0 \subseteq X$, and a set of final states $F \subseteq Q$.

The states in X (for neXt) are right-moving and the states in Y (for Yesterday) are left-moving. The tape alphabet is $\Gamma \dot\cup \{\triangleright, \triangleleft\}$. The symbol $\triangleright$ is the *left-end marker* and $\triangleleft$ is the *right-end marker*. We write $z \xrightarrow{a} z'$ instead of $(z, a, z') \in \delta$.

On input $u = a_1 \cdots a_n \in \Gamma^*$, the tape content is $\triangleright\, u \,\triangleleft$. Position 0 is labeled by $\triangleright$ and position $n+1$ has label $\triangleleft$. A transition $(z,i) \vdash (z',j)$ between configurations in $Q \times \mathbb{N}$ exists if $z \xrightarrow{a_i} z'$ and if $z' \in X$, then $j = i+1$; otherwise $j = i-1$. A computation is a sequence of transitions $(z_0, i_0) \vdash \cdots \vdash (z_t, i_t)$ with $z_0 \in X_0$, $i_0 = 1$ and $i_t = n+1$. It is accepting if z_t is a final state. Note that the left-end marker $\triangleright$ cannot be overrun by $\mathcal{A}$. A word $u \in \Gamma^*$ is accepted by $\mathcal{A}$ if there exists an accepting computation of $\mathcal{A}$ on input u. The language *recognized* by $\mathcal{A}$ is $L(\mathcal{A}) = \{u \in \Gamma^* \mid \mathcal{A} \text{ accepts } u\}$.

As usual, the automaton $\mathcal{A}$ is *deterministic* if $|X_0| = 1$ and for all $z \in Q$ and all $a \in \Gamma \cup \{\triangleright\}$ there is at most one $z' \in Q$ with $z \xrightarrow{a} z'$. It is *complete* if for all $z \in Q$ and all $a \in \Gamma$ there is at least one $z' \in Q$ with $z \xrightarrow{a} z'$ and, in addition, for all $z \in Y$ there is at least one $z' \in X$ with $z \xrightarrow{\triangleright} z'$. Every automaton can be made complete by adding a new sink state to X. The automaton is *partially ordered* (or *very weak*, or a *one-pass po2-automaton*) if there exists a partial ordering $\sqsubseteq$ of the states such that transitions are non-descending, i.e., if $z \xrightarrow{a} z'$, then $z \sqsubseteq z'$. In a partially ordered automaton, once a state is left, it is never re-entered.

Usually the disjunction of two-way automata $\mathcal{A}$ and $\mathcal{B}$ is done by simulating $\mathcal{A}$ and if this automaton rejects, then the computation restarts with a simulation of $\mathcal{B}$. This is not possible for one-pass automata since the computation stops as soon as the right-end marker is encountered. By a product automaton construction for one-pass po2-automata this problem can be solved leading to the following proposition. Similar techniques were used for deterministic po2-Büchi automata [8].

Proposition 3. *Deterministic one-pass po2-automata are closed under Boolean operations.*

Proof. Let $L \subseteq \Gamma^*$ be recognized by the complete deterministic one-pass po2-automaton $\mathcal{A} = (Q, \Gamma, \delta, x_0, F)$. The complement $\Gamma^* \setminus L$ is recognized by the automaton $\overline{\mathcal{A}} = (Q, \Gamma, \delta, x_0, Q \setminus F)$ obtained from $\mathcal{A}$ by complementing the set of final states.

For closure under union and intersection we give a product automaton construction. The main problem in this construction comes from the case when the automata do not agree on the direction in which the input is processed. To overcome this, we have to do some additional book-keeping. We only give a high-level description of the construction; the details are similar to the situation for deterministic po2-Büchi automata [8].

We simulate the two automata in parallel in what we call the *synchronous mode* as long as both automata are moving to the right. If at least one of the automata is moving to the left, then we start a simulation of one left-moving automaton in *asynchronous mode* while suspending the other automaton. We refer to the position of the input, where this disagreement on moving to the right happens, as the *synchronization point*. In asynchronous mode, the active automaton can move in either direction. As soon as the synchronization point is reached again and both automata agree on moving to the right, we switch back to the synchronous mode and we continue simulating both automata in parallel; otherwise we stay in asynchronous mode while simulating one of the automata. However, in order to apply this idea we must be able to recognize the synchronization point while computing in the asynchronous mode.

For re-synchronization we exploit the following combinatorial property of the computation of deterministic po2-automata. Assume that we are in the asynchronous mode and let $u_1 a_1 \cdots u_m a_m u'$ be such that the a_i's correspond to the positions where in the synchronous mode at least one of the automata changed its state. By determinism we have $a_i \notin \operatorname{alph}(u_i)$, and since both automata are partially ordered, m is bounded by the total number of states in both automata. In asynchronous mode, recognizing the synchronization point (the position of a_m in the above factorization) relies on keeping the following information up to date: the word $a_1 \cdots a_m$ and an index $k \in \{1, \ldots, m\}$ such that

- the current position does not lie within the prefix $u_1a_1 \cdots u_{k-1}a_{k-1}$, and
- there is a scattered subword $a_k \cdots a_m$ between the current position and the synchronization point.

Here, $a_k \cdots a_m$ is a scattered subword of w if $w \in \Gamma^* a_k \Gamma^* \cdots a_m \Gamma^*$. One can verify that we indeed can maintain this invariant for k, cf. [8, Proposition 1]. Moreover, after at most $m-1$ updates of this information there is a state change in the simulated automaton. Therefore, the simulation of the active automaton together with the above book-keeping is partially ordered. Since $a_m \notin \operatorname{alph}(u_m)$, we know that, when reading a_m in an X-state while $k = m$, we have reached the synchronization point. The exact details are slightly more technical, because we have to distinguish between X- and Y-states. □

Next, we introduce rankers. Informally a ranker is a sequence of instructions of the form "go to the next a-position" and "go to the previous a-position". More formally, a *ranker* is a word over the alphabet $\{\mathsf{X}_a, \mathsf{Y}_a \mid a \in \Gamma\}$. For a word $u \in \Gamma^*$ and a position i we define $\mathsf{X}_a(u,i) = \min\{j > i \mid u_j = a\}$ and $\mathsf{X}_a(u,i) = \max\{j < i \mid u_j = a\}$. The minimum and the maximum of the empty set are undefined. Therefore, $\mathsf{X}_a(u,i)$ is the next a-position of u after position i and $\mathsf{Y}_a(u,i)$ is the previous a-position of u before i. We extend this to rankers by setting $\mathsf{Z}_a r(u,i) = r(u, \mathsf{Z}_a(u,i))$ for $\mathsf{Z} \in \{\mathsf{X}, \mathsf{Y}\}$ and for a ranker r. In particular, rankers are processed from left to right. Rankers of the form $\mathsf{X}_a r$ are X-*rankers*. We set $\mathsf{X}_a r(u) = r(u, \mathsf{X}_a(u,0))$, i.e., X-rankers start at the left. For example, $\mathsf{X}_a\mathsf{Y}_b(bab) = 1$ whereas $\mathsf{X}_b\mathsf{Y}_a(bab)$ is undefined. The language $L(r)$ generated by an X-ranker r is the set of all words $u \in \Gamma^*$ such that r is defined on u, c.f. [3]. A language is an X-*ranker language* if it is a Boolean combination of languages generated by X-rankers.

A *monomial* is a language of the form $P = A_1^* a_1 \cdots A_k^* a_k A_{k+1}^*$ for $A_i \subseteq \Gamma$. It is *unambiguous* if for all $u \in P$ there exists a unique factorization $u = u_1 a_1 \cdots u_k a_k u_{k+1}$ with $u_i \in A_i^*$.

We are now ready to characterize the expressive power of deterministic one-pass po2-automata. Frequently, the proof of the following theorem relies on results over infinite words. This is not surprising: In some sense, Boolean combinations of Cantor sets are "concealing" the end of finite words, and infinite words do not even end at all. An essential step in the proof of the following theorem is Proposition 3, which cannot be easily deduced from the respective result over infinite words.

Theorem 9. *Let $L \subseteq \Gamma^*$. The following are equivalent:*

1. *$L \in \mathcal{DA}(\Gamma^*)$ and L is a Boolean combination of Cantor sets in Γ^*.*
2. *L is a Boolean combination of Cantor sets in $\mathcal{DA}(\Gamma^*)$.*
3. *The syntactic homomorphism $h_L : \Gamma^* \to \operatorname{Synt}(L)$ satisfies $\operatorname{Synt}(L) \in \mathbf{DA}$ and $h_L(L)$ is a union of $\mathcal{R}$-classes.*
4. *L is a finite union of unambiguous monomials $A_1^* a_1 \cdots A_k^* a_k A^*$ such that there exists no $i \in \{1, \ldots, k\}$ with $\{a_i, \ldots, a_k\} \subseteq A_i$.*
5. *L is an X-ranker language.*
6. *L is recognized by some deterministic one-pass po2-automaton.*

Proof. The equivalences "$1 \Leftrightarrow 2 \Leftrightarrow 3$" follow from Theorem 5 and Theorem 8. We say that a monomial $A_1^* a_1 \cdots A_k^* a_k A^*$ is *restricted* if there is no $i \in \{1, \ldots, k\}$ with $\{a_i, \ldots, a_k\} \subseteq A_i$.

"$2 \Rightarrow 4$": By Theorem 5, L is recognized by some $h : \Gamma^* \to M \in \mathbf{DA}$ such that $h(L)$ is a union of $\mathcal{R}$-classes. For $x \in M$ we set $[x] = h^{-1}(x)$. Let

$$K = \bigcup \left\{ [s][e]^\omega \;\middle|\; [s] \subseteq L \text{ and } s = se,\ e^2 = e \right\} \subseteq \Gamma^\infty.$$

Here, Γ^∞ is the set of all finite and infinite words over the alphabet Γ, and $[e]^\omega$ is the set of all words $u_1u_2\cdots$ with $u_i \in [e]$. Since $\varepsilon^\omega = \varepsilon$ for the empty word ε, the set $[1]^\omega$ contains also finite words. Then $K \cap \Gamma^* = L$ and K is recognized by h. Moreover, if $s = se$ and $t = tf$ for $e^2 = e$ and $f^2 = f$ with $s \mathcal{R} t$, then $[s][e]^\omega \subseteq K$ if and only if $[t][f]^\omega \subseteq K$. This is because $h(L)$ is a union of $\mathcal{R}$-classes. The language K is a finite union of restricted unambiguous monomials $A_1^*a_1 \cdots A_k^*a_kA^\infty$ over Γ^∞, see [6, Theorem 6.6]. Here, A^∞ is the set of all finite and infinite words over the alphabet A. Therefore, $L = K \cap \Gamma^*$ is a finite union of restricted unambiguous monomials $A_1^*a_1 \cdots A_k^*a_kA^*$ over Γ^*.

"4 $\Rightarrow$ 5": Let L be a finite union of restricted unambiguous monomials $A_1^*a_1 \cdots A_k^*a_kA^*$ and let $K \subseteq \Gamma^\infty$ be obtained by replacing these monomials by $A_1^*a_1 \cdots A_k^*a_kA^\infty$. Then $K \cap \Gamma^* = L$ and K is a union of restricted unambiguous monomials over Γ^∞. Now, K is definable in the first-order fragment $\Delta_2[<]$ over Γ^∞, see [6, Theorem 6.6]. Thus, K is an X-ranker language over Γ^∞, see [3, Theorem 3]. It follows that $L = K \cap \Gamma^*$ is an X-ranker language over Γ^*.

"5 $\Rightarrow$ 6": It is easy to see that every language $L(r)$ for an X-ranker r is recognizable by a deterministic one-pass po2-automaton. By Proposition 3 we get a deterministic one-pass po2-automaton for any Boolean combination of such languages.

"6 $\Rightarrow$ 1": Let L be recognized by a complete deterministic one-pass po2-automaton $\mathcal{A}$. In particular, $\mathcal{A}$ is a deterministic po2-automaton. Therefore, $\mathrm{Synt}(L) \in \mathbf{DA}$ by [16, Theorem 3.1]. Let n be any number greater than the number of states of $\mathcal{A}$, and let $x, y, z \in \Gamma^*$. We claim that $z(xy)^n \in L$ if and only if $z(xy)^nx \in L$: Consider the run of $\mathcal{A}$ on either word. Let q be the state in which $\mathcal{A}$ leaves the prefix $z(xy)^n$ for the first time. Note that this must happen eventually since $\mathcal{A}$ is complete and cannot overrun the left-end marker $\triangleright$. Then q is right-moving and moreover, q has a self-loop for all letters in $\mathrm{alph}(xy)$ by choice of n. Hence, $\mathcal{A}$ encounters the right-end marker $\triangleleft$ in the state q on both inputs $z(xy)^n$ and $z(xy)^nx$. Therefore, $z(xy)^n$ is accepted if and only if $z(xy)^nx$ is accepted. By Theorem 5, the language L is a Boolean combination of Cantor sets. □

Membership in the variety of finite monoids **DA** is decidable. Therefore, all of the above properties are decidable by Corollary 3.

Remark 3. We use the shortcuts "dfa" and "nfa" for *deterministic finite automata* and *nondeterministic finite automata*, respectively. We write po1 for *partially ordered one-way* and po2 for *partially ordered two-way*. Using this notation, we have the following inclusions between language classes recognizable by partially ordered automata:

$$\text{po1-dfa} \subsetneq \text{one-pass po2-dfa} \subsetneq \text{po2-dfa} \subsetneq \text{po2-nfa} = \text{po1-nfa}.$$

The following (very similar) languages show that the inclusions are strict. The unambiguous monomial $\{a,c\}^* ab \{a,b,c\}^*$ is recognizable by some one-pass po2-dfa, but it is not recognizable by any po1-dfa. The language $\{a,b,c\}^* ab \{b,c\}^*$ is recognizable by some po2-dfa, but it is not recognizable by any one-pass po2-dfa. Finally, the language $\{a,b,c\}^* ab \{a,b,c\}^*$ is recognizable by some po1-nfa, but it is not recognizable by any po2-dfa. The equivalence of po2-nfa and po1-nfa is due to Schwentick, Thérien, and Vollmer [16]. For each of the above language classes the membership problem is decidable: The class po1-dfa corresponds to $\mathcal{R}$-trivial monoids [16], one-pass po2-dfa correspond to $\mathcal{R}$-classes of monoids in **DA** (Theorem 5 and Theorem 9), the algebraic equivalent of po2-dfa is the variety of finite monoids **DA** [16], and po2-nfa are expressively complete for the level 3/2 of the Straubing-Thérien hierarchy [16] which is decidable [14]. □

6 Summary

A language $L \subseteq \Gamma^*$ is a *Cantor set* if $L\Gamma^* \subseteq L$, it is a *left Cantor set* if $\Gamma^*L \subseteq L$, and it is a *two-sided Cantor set* if $\Gamma^*L\Gamma^* \subseteq L$. All Cantor sets (resp. all left Cantor sets, resp. all two-sided Cantor sets) form a topology on Γ^*. Our characterizations of regular open sets and of regular Boolean combinations of open sets are summarized in Table 1. Both the automata-theoretic characterizations and the algebraic characterizations can be used for deciding the membership problem in each of the six classes of regular languages (Corollaries 1 and 3).

A *flip automaton* is an automaton with no transitions from final to non-final states. Complete flip automata recognize exactly the class of regular Cantor sets. Moreover, every complete deterministic automaton recognizing a Cantor set is a flip automaton (Theorem 1). An automaton is *weak* if every strongly connected component is either final or non-final. Deterministic weak automata recognize Boolean combinations of Cantor sets. Again, we have a strong version of the converse: *Every* deterministic automaton recognizing a Boolean combination of Cantor sets is weak (Theorem 5).

In addition, every regular language can be written as an intersection of a closed set and of a dense set (Theorem 4). This holds for any of the three Cantor topologies. For any variety of regular languages $\mathcal{V}$ and for any language $L \in \mathcal{V}(\Gamma^*)$ it is equivalent whether L is a Boolean combination of arbitrary open sets or whether L is a Boolean combination of open sets in $\mathcal{V}(\Gamma^*)$. Again, this holds for all three Cantor topologies (Theorem 8).

As a case study, we consider the variety $\mathcal{DA}$ and we show that the Boolean combinations of Cantor sets in $\mathcal{DA}$ admit natural characterizations in terms of restricted unambiguous polynomials, X-ranker languages, and one-pass partially ordered two-way automata (Theorem 9).

Regular languages	**Algebra**	**Lattice equation**	**Automata**	
Cantor sets	$\leq_{\mathcal{R}}$-order ideals	$y \to yz$	complete flip automata	Thm. 1
left Cantor sets	$\leq_{\mathcal{L}}$-order ideals	$y \to xy$		Thm. 2
two-sided Cantor sets	$\leq_{\mathcal{J}}$-order ideals	$y \to xyz$		Thm. 3
Boolean combinations of Cantor sets	$\mathcal{R}$-classes	$z(xy)^\omega \to z(xy)^\omega x$	deterministic weak automata	Thm. 5
Boolean combinations of left Cantor sets	$\mathcal{L}$-classes	$(ts)^\omega z \to s(ts)^\omega z$		Thm. 6
Boolean combinations of two-sided Cantor sets	$\mathcal{J}$-classes	$(ts)^\omega z(xy)^\omega \to s(ts)^\omega z(xy)^\omega x$		Thm. 7

Table 1: Characterizations of regular languages with respect to Cantor topologies.

References

[1] Alpern, B., Schneider, F.B.: Defining liveness. Inf. Process. Lett. 21, 181–185 (1985)

[2] Chang, E., Manna, Z., Pnueli, A.: The safety-progress classification. In: *Logic and Algebra of Specifications*, Proc. vol. 79, pp. 143–202. Springer (1993)

[3] Dartois, L., Kufleitner, M., Lauser, A.: Rankers over infinite words. In: DLT'10, Proc. LNCS, vol. 6224, pp. 148–159. Springer (2010)

[4] Dax, C., Klaedtke, F.: Alternation elimination by complementation (Extended abstract). In: LPAR'08, Proc. LNCS, vol. 5330, pp. 214–229. Springer (2008)

[5] Diekert, V., Gastin, P., Kufleitner, M.: A survey on small fragments of first-order logic over finite words. Int. J. Found. Comput. Sci. 19(3), 513–548, special issue DLT'07 (2008)

[6] Diekert, V., Kufleitner, M.: Fragments of first-order logic over infinite words. Theory Comput. Syst. 48, 486–516 (2011)

[7] Gehrke, M., Grigorieff, S., Pin, J.É.: Duality and equational theory of regular languages. In: ICALP'08, Proc. LNCS, vol. 5126, pp. 246–257. Springer (2008)

[8] Kufleitner, M., Lauser, A.: Partially ordered two-way Büchi automata. In: CIAA'10, Proc. LNCS, vol. 6482, pp. 181–190. Springer (2011).

[9] Kupferman, O., Vardi, M.Y.: Weak alternating automata and tree automata emptiness. STOC'98, Proc. pp. 224–233. ACM Press (1998)

[10] Kupferman, O., Vardi, M.Y.: Weak alternating automata are not that weak. ACM Trans. Comput. Log. 2(3), 408–429 (2001)

[11] Muller, D.E., Saoudi, A., Schupp, P.E.: Alternating automata, the weak monadic theory of the tree, and its complexity. In: ICALP'86, Proc. LNCS, vol. 226, pp. 275–283. Springer (1986)

[12] Muller, D.E., Saoudi, A., Schupp, P.E.: Weak alternating automata give a simple explanation of why most temporal and dynamic logics are decidable in exponential time. In: LICS'88, Proc. pp. 422–427 (1988)

[13] Perrin, D., Pin, J.É.: Infinite words, Pure and Applied Mathematics, vol. 141. Elsevier (2004)

[14] Pin, J.É., Weil, P.: Polynomial closure and unambiguous product. Theory Comput. Syst. 30(4), 383–422 (1997)

[15] Schützenberger, M.P.: Sur le produit de concaténation non ambigu. Semigroup Forum 13, 47–75 (1976)

[16] Schwentick, Th., Thérien, D., Vollmer, H.: Partially-ordered two-way automata: A new characterization of DA. In: DLT'01, Proc. LNCS, vol. 2295, pp. 239–250. Springer (2001)

[17] Staiger, L.: ω-languages. In: Handbook of Formal Languages, vol. 3, pp. 339–387. Springer (1997)

[18] Tesson, P., Thérien, D.: Diamonds are Forever: The Variety DA. In: Semigroups, Algorithms, Automata and Languages, 2001. pp. 475–500. World Scientific (2002)

[19] Thomas, W.: Automata on infinite objects. In: Handbook of Theoretical Computer Science, chap. 4, pp. 133–191. Elsevier (1990)

[20] Weis, Ph., Immerman, N.: Structure theorem and strict alternation hierarchy for FO^2 on words. Log. Methods Comput. Sci. 5 (3:3), 1–23 (2009)