Graph type conditions on automata determining varieties of languages

Ondřej Klíma and Libor Polák

Department of Mathematics and Statistics
Masaryk University, Brno
Czech Republic

Semigroups and Automata, Lisboa 2016

I. Algebraic Theory of Regular Languages

Examples

- Goal of the study: effective characterizations of certain natural classes of regular languages.
- Typical result: a language belongs to a given class iff its syntactic monoid belongs to a certain class of monoids.

Theorem (Schützenberger – 1966)

A regular language L is star-free if and only if its syntactic monoid is aperiodic.

Theorem (Simon — 1972)

A regular language L is piecewise testable if and only if the syntactic monoid of L is $\mathcal{J}$-trivial.

- General framework – Eilenberg correspondence.

Varieties of Languages

Definition

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

- $\mathcal{V}(A)$ is closed under finite unions, finite intersections and complements (in particular $\emptyset, A^* \in \mathcal{V}(A)$),
- $\mathcal{V}(A)$ is closed under quotients, 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 $f^{-1}(L) = \{ v \in B^* \mid f(v) \in L \} \in \mathcal{V}(B)$.

A Formal Definition of a DFA

Definition

A deterministic finite automaton over the alphabet A is a five-tuple $\mathcal{A} = (Q, A, \cdot, i, F)$, where

- Q is a nonempty set of states,
- $\cdot : Q \times A \to Q$ is a complete transition function, which can be extended to a mapping $\cdot : Q \times A^* \to Q$ by $q \cdot \lambda = q,\ q \cdot (ua) = (q \cdot u) \cdot a$,
- $i \in Q$ is the initial state,
- $F \subseteq Q$ is the set of final states.

The automaton $\mathcal{A}$ accepts a word $u \in A^*$ iff $i \cdot u \in F$. The automaton $\mathcal{A}$ recognizes the language $L_{\mathcal{A}} = \{u \in A^* \mid i \cdot u \in F\}$.

Motivations for a Notion of a Variety of Automata

- Why monoids instead of automata?
  - An equational description of pseudovarieties of monoids by pseudoidentities.
  - Other algebraic constructions, e.g. products (semidirect, wreath, Mal'cev).
- Why are we still interested in automata characterizations?
  - Usually, a regular language is given by an automaton. And computation of the syntactic monoid need not to be effective (can be exponentially larger).
  - Sometimes a "graph condition" on automata can be easier to test than an equational condition on monoids.

So, basically there are three worlds: classes of languages, classes of (enriched) semiautomata (no initial and no final states) and those of appropriate algebraic structures.

Generalizations of the Eilenberg Correspondence

Since not all natural classes of regular languages are varieties, one of the recent directions of the research in algebraic theory of regular languages is devoted to generalizations of the Eilenberg correspondence.

- Pin (1995): Positive varieties of regular languages — closure under complementation is not required. Algebraic counterparts are pseudovarieties of finite ordered monoids.
  (Syntactic monoid is implicitly ordered.)
- Polák (1999): Conjunctive (and disjunctive) varieties.
- Straubing (2002): $\mathcal{C}$-varieties of languages.
- Ésik, Larsen (2003): literal varieties of languages.
- Gehrke, Grigorieff, Pin (2008): Lattices of regular languages.

II. Varieties of Automata

The Construction of a Minimal DFA by Brzozowski

- For a language $L \subseteq A^*$ and $u \in A^*$, we define a left quotient $u^{-1}L = \{ w \in A^* \mid uw \in L \}$.

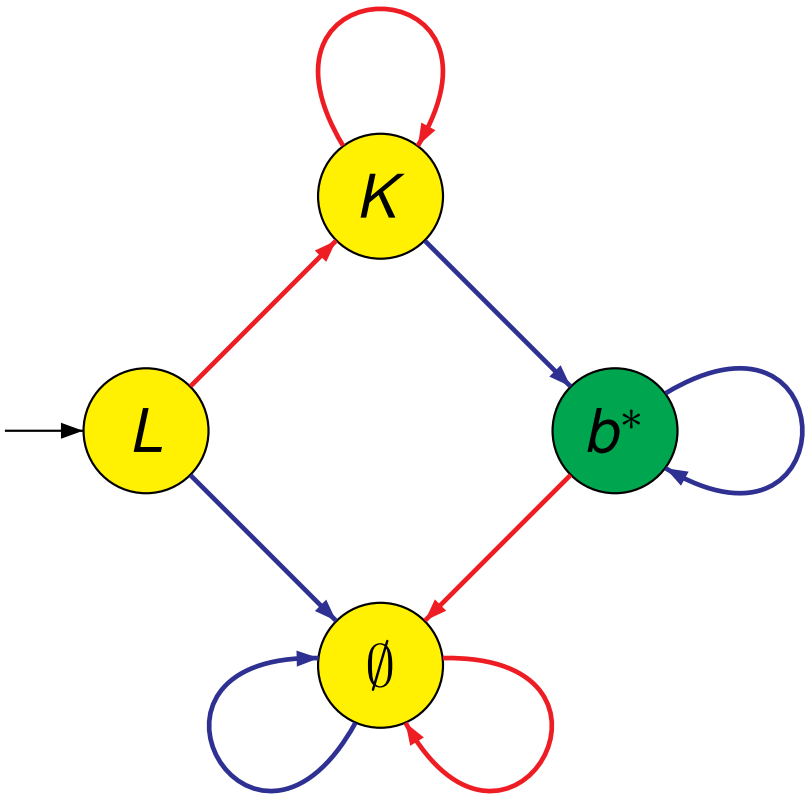
Definition

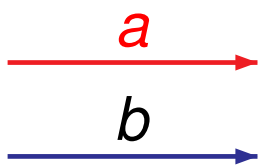
The canonical deterministic automaton of L is $\mathcal{D}_L = (D_L, A, \cdot, L, F)$, where

- $D_L = \{ u^{-1}L \mid u \in A^* \}$,
- $q \cdot a = a^{-1}q$, for each $q \in D_L$, $a \in A$,
- $q \in F$ iff $\lambda \in q$.

- Each state $q = u^{-1}L$ is formed by all words transforming the state q into a final state.

An Example of a Canonical Automaton





$$L = a^+ b^+$$

$$K = a^{-1} L = a^* b^+$$

$$b^{-1} K = b^*$$

Preimages in Morphisms, Varieties of Automata

- Let $f : B^* \to A^*$ be a morphism, We say that $(P, B, \circ)$ is an f-subautomaton of $(Q, A, \cdot)$ if $P \subseteq Q$ and $q \circ b = q \cdot f(b)$ for every $q \in P$, $b \in B$.

Definition

A variety of semiautomata $\mathbb{V}$ associates to every non-empty finite alphabet A a class $\mathbb{V}(A)$ of semiautomata (no initial nor final states) over alphabet A in such a way that

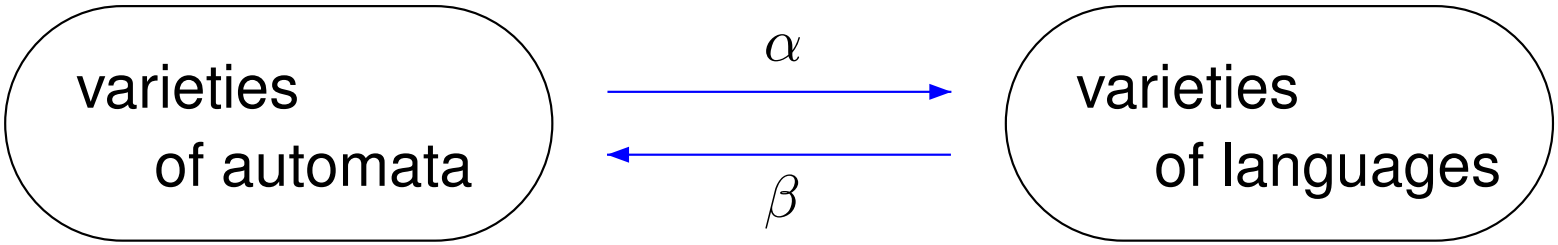
- $\mathbb{V}(A) \neq \emptyset$ is closed under disjoint unions, finite direct products and morphic images,
- $\mathbb{V}$ is closed under f-subautomata.

An Eilenberg Type Correspondence

- For each variety of semiautomata $\mathbb{V}$ we denote by $\alpha(\mathbb{V})$ the variety of regular languages given by

$$(\alpha(\mathbb{V}))(A) = \{L \subseteq A^* \mid \exists \mathcal{A} = (Q, A, \cdot, i, F) : L = L_{\mathcal{A}} \wedge (Q, A, \cdot) \in \mathbb{V}(A)\}.$$

- For each variety of regular languages $\mathcal{L}$ we denote by $\beta(\mathcal{L})$ the variety of automata generated by all DFAs $\mathcal{D}_L$, where $L \in \mathcal{L}(A)$ for some alphabet A.



Theorem (Ésik and Ito, Chaubard, Pin and Straubing)

The mappings α and β are mutually inverse isomorphisms between the lattice of all varieties of automata and the lattice of all varieties of regular languages.

- A version for $\mathcal{C}$-varieties is obvious: we consider f-subautomata (etc.) just for $f \in \mathcal{C}$.
- Ésik and Ito were working with literal varieties (morphisms map letters to letters, i.e $f(B) \subseteq A$) and used disjoint union.
- Chaubard, Pin and Straubing called the automata $\mathcal{C}$-actions and used trivial automata.

An Example – Acyclic Automata

- One of the conditions in Simon's characterization of piecewise testable languages is that a minimal DFA is acyclic.
- A content $\mathrm{c}(u)$ of a word $u \in A^*$ is the set of all letters occurring in u.
- We say that $(Q, A, \cdot)$ is a acyclic if for each $u \in A^*$ and $q \in Q$ we have

$$q \cdot u = q \implies (\forall a \in \mathrm{c}(u) : q \cdot a = q)\,.$$

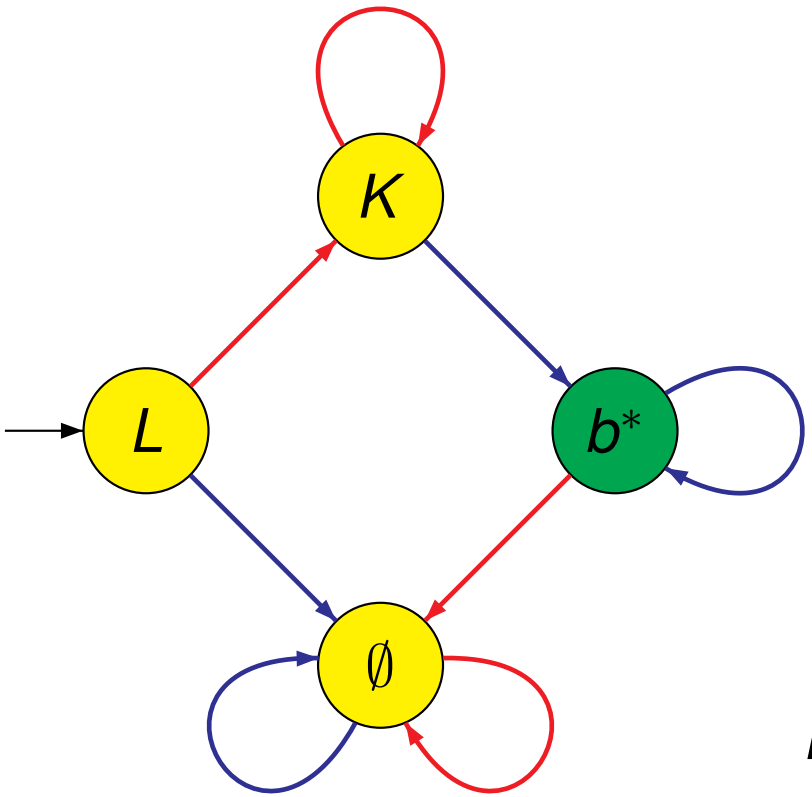
- The class of all acyclic semiautomata is a variety.
- The corresponding variety of languages (well-known): (disjoint) unions of the languages of the form

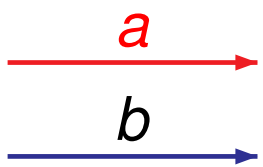
$$A_0^* a_1 A_1^* a_2 A_2^* \dots A_{n-1}^* a_n A_n^*, \quad \text{where} \quad a_i \notin A_{i-1} \subseteq A\,.$$

An Example – Piecewise Testable Languages

- In DLT'13 we gave an alternative condition for automata recognizing piecewise testable languages.
- We call an acyclic semiautomaton $(Q, A, \cdot)$ locally confluent, if for each state $q \in Q$ and every pair of letters $a, b \in A$, there is a word $w \in \{a, b\}^*$ such that $(q \cdot a) \cdot w = (q \cdot b) \cdot w$.
- A stronger condition: an acyclic semiautomaton $(Q, A, \cdot)$ is confluent, if for each state $q \in Q$ and every pair of words $u, v, \in \{a, b\}^*$, there is a word $w \in \{a, b\}^*$ such that $(q \cdot u) \cdot w = (q \cdot v) \cdot w$.
- Each acyclic semiautomaton is confluent iff it is locally confluent.
- The class of all acyclic confluent semiautomata is a variety which corresponds to the variety of p. t. languages.

An Example of a Piecewise Testable Language





$$L = a^+b^+$$

$$K = a^{-1}L = a^*b^+$$

$$b^{-1}K = b^*$$

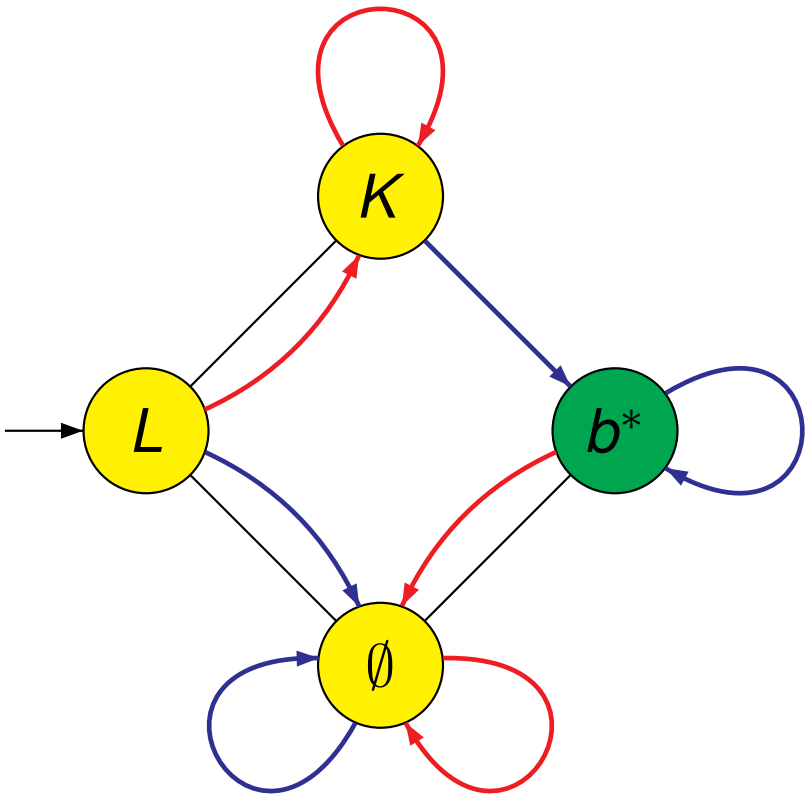
$$L = A^*aA^*bA^* \cap (A^*bA^*aA^*)^c$$

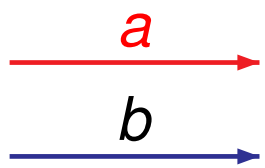
III. Automata Enriched with an Algebraic Structure

A Natural Ordering of the Canonical Automaton

- For a language $L \subseteq A^*$, we have defined a the canonical deterministic automaton: $\mathcal{D}_L = (D_L, A, \cdot, L, F)$, where
  - $D_L = \{ u^{-1}L \mid u \in A^* \}$,
  - $q \cdot a = a^{-1}q$, for each $q \in D_L$, $a \in A$,
  - $q \in F$ iff $\lambda \in q$.
- Therefore states are ordered by inclusion, which means that each minimal automaton is implicitly equipped with a partial order.
- The action by each letter a is an isotone mapping: for all states p, q such that $p \subseteq q$ we have $p \cdot a = a^{-1}p \subseteq a^{-1}q = q \cdot a$.
- The final states form an upward closed subset w.r.t. $\subseteq$.

An Example of an Ordered Automaton



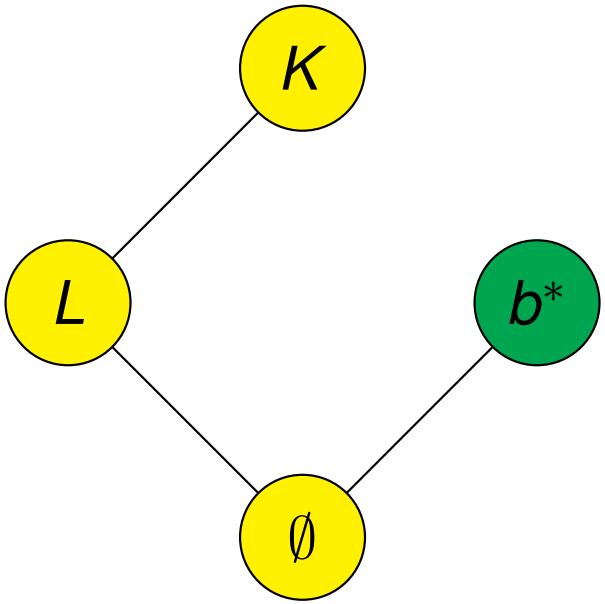


$L = a^+b^+$

$K = a^{-1}L = a^*b^+$

$L \subseteq K$

An Example of an Ordered Automaton



$$L = a^+ b^+$$

$$K = a^{-1} L = a^* b^+$$

$$L \subseteq K$$

An Ordered Automaton

Definition

An ordered automaton over the alphabet A is a six-tuple $\mathcal{A} = (Q, A, \cdot, \leq, i, F)$, where

- $\mathcal{A} = (Q, A, \cdot, i, F)$ is a usual DFA;
- $\leq$ is a partial order;
- an action by every letter is an isotone mapping from the partial ordered set $(Q, \leq)$ to itself;
- F is an upward closed set, i.e. $p \leq q, p \in F$ implies $q \in F$.

An Eilenberg Type Correspondence

Definition

A variety of ordered semiautomata $\mathbb{V}$ associates to every non-empty finite alphabet A a class $\mathbb{V}(A)$ of ordered semiautomata over alphabet A in such a way that

- $\mathbb{V}(A) \neq \emptyset$ is closed under disjoint union, finite direct products and morphic images,
- $\mathbb{V}$ is closed under f-subautomata.

Theorem (Pin)

There are mutually inverse isomorphisms between the lattice of all varieties of ordered semiautomata and the lattice of all positive varieties of regular languages.

The Level 1/2

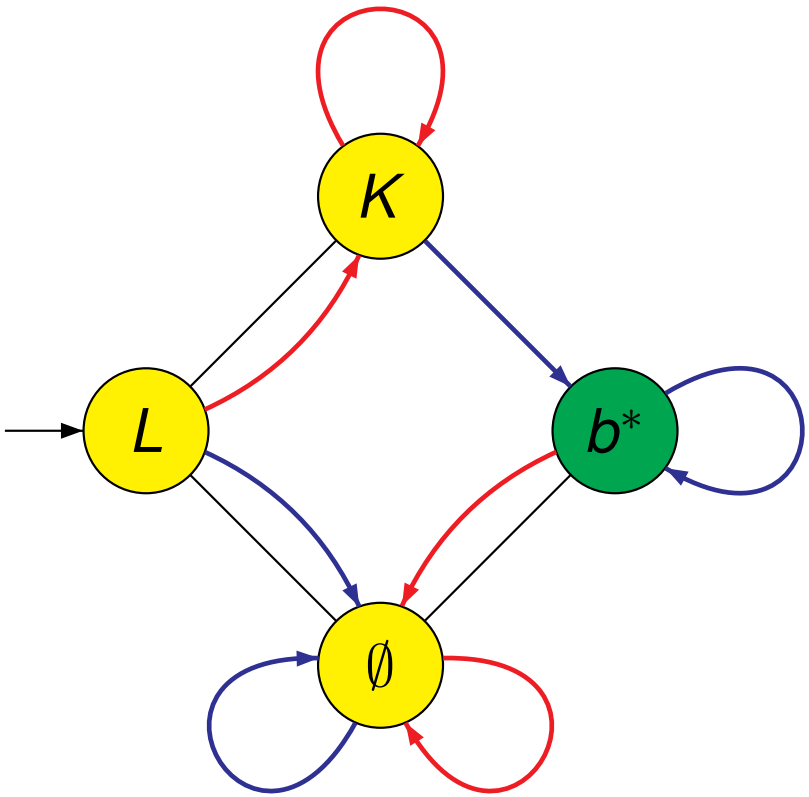
- Piecewise testable languages are Boolean combinations of languages of the form

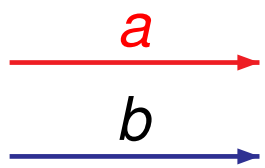
$$A^* a_1 A^* a_2 A^* \dots A^* a_\ell A^*, \text{ where } a_1, \dots, a_\ell \in A, \ \ell \geq 0 .$$

- Piecewise testable languages form level 1 in Straubing-Thérien hierarchy.
- Level 1/2 is formed just by finite unions of intersections of languages above.
- The corresponding variety of ordered semiautomata is the class of all ordered semiautomata where actions by letters are non-decreasing mappings. I.e. ordered automata satisfying:

$$\forall q \in Q, a \in A : q \cdot a \geq q .$$

An Example of an Ordered Automaton outside 1/2





$L = a^+b^+$

$K = a^{-1}L = a^*b^+$

$L \not\subseteq L \cdot b = \emptyset$

IV. Presentations of classes of languages via automata

Using forbidden patterns

Following Iván, a pattern is a triple $\mathcal{P} = (V, E, \ell)$ where (V, E) is a finite oriented graph and ℓ labels the edges by variables from the set X. A semiautomaton $\mathcal{A} = (Q, A, \cdot)$ admits $\mathcal{P}$ if there exists an injective mapping $f : V \to Q$ and a mapping $h : X \to A^+$ such that: for each $(k, l) \in E$, we have $f(k) \cdot h(\ell(k, l)) = f(l)$. Otherwise, $\mathcal{A}$ avoids $\mathcal{P}$.

Examples:

- $\mathcal{P}_f = (\{k, l\}, \{(k, k), (l, l)\}, \ell)$, $\ell(k, k) = x, \ell(l, l) = y$.
- $\mathcal{P}_d = (\{k, l\}, \{(k, k), (l, l)\}, \ell)$, $\ell(k, k) = \ell(l, l) = x$.
- $\mathcal{P}_r = (\{k, l\}, \{(k, k), (k, l)\}, \ell)$, $\ell(k, k) = x, \ell(k, l) = y$.
- $\mathcal{P}_g = (\{k, l\}, \{(k, k), (k, l), (l, l)\}, \ell)$, $\ell(k, k) = x$, $\ell(k, l) = y, \ell(l, l) = x$.

Using forbidden patterns II

Iván shows that the languages for which the minimal complete DFAs avoids those patterns are exactly finite or cofinite, definite, reverse definite and generalized definite languages, respectively. (Some results already known before.)

In several Pin's papers one can find an another kind of conditions:
In a DFA $\mathcal{A} = (Q, A, \cdot, i, F)$ there are

- (1) no $p, q, r \in Q$, $p \neq q \neq r$ for which there are $u, v \in A^*$ with $p \cdot u = q \cdot u = q, q \cdot v = r$,
- (2) no $p, q, r, s, t \in Q$, $p \neq t$ for which there are $u, v \in A^*$ such that
  $p \cdot u = q \cdot u = p, q \cdot v = r \cdot v = q, r \cdot u = s \cdot u = s, s \cdot v = t \cdot v = t$,

Some injectivity conditions

- (3) no $p, q, r, s \in Q, p \notin F, s \in F$ for which there are $u, v \in A^*$ such that $q \cdot v = p, q \cdot u = r \cdot u = r, r \cdot v = s$,
- (4) no $p, q, r \in Q,\ q \neq r$ for which there are $u, v \in A^*$ with $p \cdot v = p, p \cdot u = q \cdot u = q, q \cdot v = r \cdot v = r$.

A language L over A is reversible if it is accepted by a NFA $\mathcal{A} = (Q, A, E, I, F),\ E \subseteq Q \times A \times Q$, not necessarily complete, nor necessarily $| I | = 1$ such that the action of each $a \in A$ on Q is both deterministic a codeterministic.
L is bideterministic if moreover $\mathcal{A}$ can be taken with $| I | = | F | = 1$.

Using forbidden patterns III

We denote the above classes by $\mathcal{R}$ and $\mathcal{BD}$. Deciding a membership of a given L in $\mathcal{BD}$ is easy - one looks at the minimal complete DFA.

Pin also observed that (2) and (3) in minimal complete DFA characterize the class $\mathcal{R}$ from the last subsection. Also patterns (1) and (2), respectively, characterize classes of languages close to reversible ones.

General theory

Our pattern is a triple $\mathcal{P} = (V, E, \ell)$ where (V, E) is a finite oriented graph where multiple loops are allowed and ℓ labels the edges by variables from the set X. Also V is equipped with relations $\neq$ and $\not\leq$ and X by c (the same content) and $\not\imath$ (first letters different). An ordered semiautomaton $\mathcal{A} = (Q, A, \cdot, \leq)$ admits $\mathcal{P}$ if there exists a mapping $f : V \to Q$ respecting $\neq$ and $\not\leq$ and a mapping $h : X \to A^+$ respecting c and $\not\imath$, some x's allowed to go to A^*, such that: for each $(k, l) \in E$, we have $f(k) \cdot h(\ell(k, l)) = f(l)$. Otherwise, $\mathcal{A}$ avoids $\mathcal{P}$.

General theory II

Moreover, one can consider a category $\mathcal{C}$ of finitely generated free monoids and to ask that all $(h : X^* \to A^*)$'s are from $\mathcal{C}$. One can assign to each pattern $\mathcal{P}$ and a category $\mathcal{C}$ the class $\mathbb{V}(\mathcal{P}, \mathcal{C})$ of ordered semiautomata avoiding $\mathcal{P}$ with respect to $\mathcal{C}$. Notice that the languages given by $\mathbb{V}(\mathcal{P}, \mathcal{C})$ and morphic images of $\mathbb{V}(\mathcal{P}, \mathcal{C})$ are the same.

Nasty example 1 $\mathbb{V}(\mathcal{P}, \mathcal{C})$ is not closed with respect to morphic images.
Nasty example 2 $\mathbb{V}(\mathcal{P}, \mathcal{C})$ is not closed with respect to the finite products.

Open problems

Problems

- Which positive varieties of languages we can characterize using forbidden patterns? (Of course, all classes characterizable by an identity or an inequality for corresponding syntactic structures.)
- How to deal with the characterizations using minimal trim DFA? (sparse languages are OK)
- Characterize configurations for which the corresponding class of ordered semiautomata is closed with respect to morphic images.