Open problems about regular languages, 35 years later

Jean-Éric Pin

LIAFA, CNRS and University Paris Diderot

June 24, 2015, Brzozowski 80

35 years ago...

In 1980, Janusz A. Brzozowski presented a selection of six open problems about regular languages and mentioned two other problems in the conclusion of his article.

These problems have been the source of some of the greatest breakthroughs in automata theory over the past 35 years.

What is known on these questions and what are the hopes for the next 35 years?

Summary

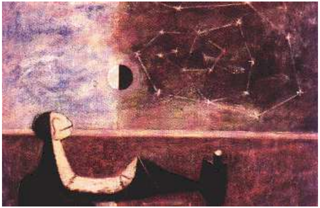
(1) Star height
(2) Restricted star height
(3) Group complexity
(4) Star removal
(5) Regularity of non-counting classes
(6) Optimality of prefix codes

Bonus...

(1) Limitedness problem
(2) Dot-depth hierarchy

Part I

Star height problems



El Hombre ante el Infinito
Rufino Tamayo (1950)

- Limitedness problem
- Restricted star height
- Star height

Limitedness problem

Problem (1966). Given a regular language L, can one decide whether there exists an integer n such that $L^n = L^*$.

Preliminary work: Linna (1973).

Solutions by Hashiguchi (1978), Simon (1978), short semigroup solution by Kirsten (2002).

Theorem (Leung and V. Podolskiy 2004, Kirsten 2004)

The limitedness problem is PSPACE-complete.

The finite range problem for weighted automata.

Decide whether or not the behaviour of a given weighted automaton has a finite range.

Theorem (Mandel and Simon 1977)

The finite range problem for $(\mathbb{N}, +, \times)$ is decidable.

Theorem (Hashiguchi 1982, 1986, Leung 1987, Simon 1978–1994)

The finite range problem for the tropical semiring $(\mathbb{N} \cup \{+\infty\}, \min, +)$ is decidable.

Restricted star height

A regular expression of height 3

$$\left(\left(a(ba)^*b\right)^* + \left(b(aa)^*b + c\right)^*\right)^*$$

Problem (Eggan 1963): compute the minimal star height of a regular expression representing a given regular language.

Dejean and Schützenberger (1966): for each $n \geqslant 0$, there exists a language of star height n.

Restricted star height

- Hashiguchi (1982): Star height one is decidable.
- Hashiguchi (1988): Star height is decidable.
- Kirsten (2005): Simplified proof. Complexity in double exponential space.
- Fijalkow, Gimbert, Kelmendi, Kuperberg: first implementation. http://www.liafa.univ-paris-diderot.fr/~nath/?page=acmepp

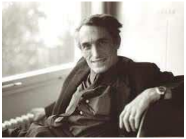
Impact of the restricted star height problem.

- Simon: Max-plus and tropical semirings
- Colcombet: Regular cost functions, stabilisation monoids

Star height

Same question, but complementation is allowed.

$$\begin{aligned} A^* &= \emptyset^c \\ (ab)^* &= \left(bA^* \cup A^*a \cup A^*aaA^* \cup A^*bbA^*\right)^c \\ &= \left(b\emptyset^c \cup \emptyset^c a \cup \emptyset^c aa\emptyset^c \cup \emptyset^c bb\emptyset^c\right)^c \end{aligned}$$



Schützenberger (1965):
algebraic characterization
of star-free languages.

The language $(aa)^*$ has star height 1, but no language of star height > 1 is known!

Star height

Theorem (Pin 1978, Brzozowski 1980)

If the languages of star-height $\leqslant 1$ form a variety of languages, then all regular languages have star-height $\leqslant 1$.

Theorem (Pin, Straubing, Thérien 1989)

For each n, the class of all languages of star-height $\leqslant n$ is closed under Boolean operations, residuals and inverse of length-preserving morphisms.

Hopes

Looking for a language of star-height $\geqslant 2$? Take any monoid morphism $\pi : A^* \to G$, where G is any complicated group and take $L = \pi^{-1}(1)$.

The languages of star-height $\leqslant n$ form a length-preserving variety of languages and hence can be defined by length-preserving profinite equations.

It would suffice to find a single nontrivial equation satisfied by all languages of star-height $\leqslant 1$ to prove the existence of a language of star-height > 1.

Other suggestions

What is the length-preserving variety of languages generated by the languages F^*, where F is finite?

Daviaud and Paperman (MFCS 2015) found equations characterizing the closure under Boolean operations and residuals of the set $\{u^* \mid u \text{ is a word}\}$.

Intermediate star-height: just allow union and intersection but no complement in the definition of the star height. Are there languages of arbitrary intermediate star-height? Is the intermediate star-height decidable?

Group complexity

Theorem (Krohn-Rhodes 1966)

Every finite semigroup S divides a wreath product of the form

$$A_0 \circ G_1 \circ A_1 \cdots A_{n-1} \circ G_n \circ A_n \qquad (*)$$

where $A_0, A_1, \ldots, A_n$ are aperiodic semigroups and $G_1, \ldots, G_n$ are groups.

The group complexity of S is the smallest possible integer n over all decompositions of type $(*)$.

The complexity problem for finite semigroups

Problem (Rhodes). Is there an algorithm to compute the complexity of a given finite semigroup?

Theorem (Karnofsky-Rhodes, 1982)

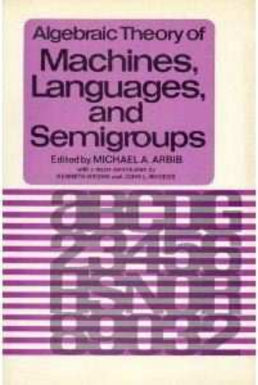
One can decide whether a finite semigroup divides a wreath product of the form $G \circ A$.

Theorem (Karnofsky-Rhodes, 1982)

One can decide whether a finite semigroup divides a wreath product of the form $A \circ G$.

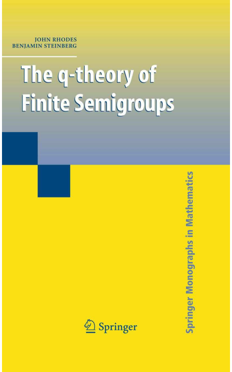
Books

1968



Arbib (ed.)

2011



Rhodes and Steinberg

Impact and hopes

Important notions in semigroup theory: Rhodes expansion, pointlike sets, relational morphisms, ...

However, major progress came from language theory. The characterization of locally testable languages (Brzozowski-Simon, McNaughton), of dot-depth one languages (Knast), the variety theorem (Eilenberg), the wreath product principle (Straubing), lead to new ideas in the study of the wreath product.

Best hope for the solution of the complexity problem: Ben Steinberg.

Star removal

Let K be a regular language. Then the equation $K = XK$ has a maximal solution L^*. Then $K = L^*K$ and the equation in R

$$K = L^*R$$

has a minimal solution $R = K - (L^* - 1)K$.

Iterating this process on R, we get a decomposition

$$K = L_1^* L_2^* \cdots L_k^* R_k$$

where R_k is the minimal solution of $K = L_1^* L_2^* \cdots L_k^* R$. Does this process terminates (i.e. $L_k^* = 1$ at some point)?

Regularity of non-counting classes

Let $\sim_n$ be the smallest congruence on A^* satisfying $x^n \sim_n x^{n+1}$ for all $x \in A^*$ and let $\mu : A^* \to A^*/{\sim_n}$ be the natural morphism.

Problem. Is $\mu^{-1}(m)$ a regular language for every $m \in A^*/{\sim_n}$?

Extended version (McCammond). Let $\sim_{n,m}$ be the congruence on A^* generated by the relations $x^n \sim x^{n+m}$. Are the congruence classes regular?

Regularity of non-counting classes

Theorem (de Luca-Varricchio 90, McCammond 91, Guba 93, Do Lago 96-98)

The conjecture holds for $n \geqslant 3$ *and any* $m > 0$.

Theorem (Do Lago 96)

The conjecture does not hold for $n = 2$ *and* $m > 1$.

Regularity of non-counting classes

For $n = 2$, $m = 1$ ($x^3 = x^2$), the problem is still open.

Theorem (Plyushchenko and Shur 2011)

For $n = 2$ and $m = 1$, the conjecture holds for all the elements containing an overlap-free or an almost overlap-free word.

Impact

Related to the Burnside problem.

- Is a k-generated group satisfying the identity $x^n = 1$ necessarily finite?
- $B(k,3)$, $B(k,4)$, and $B(k,6)$ are finite for all k.
- The case $B(2,5)$ is still open.

Optimality of prefix codes

A subset X of A^+ is a code if the condition

$$x_1 \cdots x_n = x'_1 \cdots x'_m \quad \text{(where } x_i, x'_i \in X\text{)}$$

implies $n = m$ and $x_i = x'_i$ for $i = 1, \ldots, n$. It is a prefix code if any two distinct words in X are incomparable for the prefix order.

Let $\alpha : A^* \to \mathbb{N}^A$ be defined by $\alpha(u) = (|u|_a)_{a \in A}$. Extended to series. If $X = ba + abab + baab + bbab$, then $\alpha(X) = ab + 2a^2b^2 + ab^3$.

A language X is commutatively prefix if $\alpha(X) = \alpha(P)$ for some prefix code P.

Conjecture 1 [Schützenberger (1956)]. Every code is commutatively prefix.

Couterexample [P. Shor (1983)].
$X = \{ba, ba^7, ba^{13}, ba^{14}, a^3b, a^3ba^2, a^3ba^4, a^3ba^6,$
$a^8b, a^8ba^2, a^8ba^4, a^8ba^6, a^{11}b, a^{11}ba^2, a^{11}ba^4\}$.

Conjecture 2. Every finite maximal code is commutatively prefix.

Theorem [14.6.4] A set X is commutatively prefix iff the series $(1 - \alpha(X))/(1 - \alpha(A))$ has nonnegative coefficients.

The factorization conjecture

Factorization Conjecture. For any finite maximal code X over A, there exist two polynomials $P, S \in \mathbb{N}\langle A\rangle$ such that $1 - X = P(1 - A)S$.

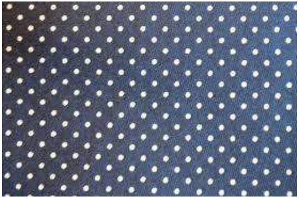
Theorem. The factorization conjecture implies that every finite maximal code is commutatively prefix.

Related to Kraft's inequality.

Theorem [Reutenauer (1985)]. For any finite maximal code X over A, there exist polynomials $P, S \in \mathbb{Z}\langle A\rangle$ such that $1 - X = P(1 - A)S$.

Part II

Dot depth hierarchy



- Operations on regular languages
- Dot-depth hierarchy
- Connection with logic

Operations on regular languages

Let $\mathcal{L}$ be a class of languages.

- $\mathcal{BL}$ is the Boolean closure of $\mathcal{L}$.
- $\mathsf{Pol}(\mathcal{L})$ is the polynomial closure of $\mathcal{L}$: unions of products of the form $L_0a_1L_1a_2 \cdots a_kL_k$ where $L_0, \ldots, L_k$ are in $\mathcal{L}$ and $a_1, \ldots, a_k$ are letters.
- $\mathsf{UPol}(\mathcal{L})$ is the unambiguous polynomial closure of $\mathcal{L}$: unions of unambiguous products

Dot-depth hierarchy

Let $\mathcal{B}_0$ be the class of finite/cofinite languages.
Let $\mathcal{B}_n = (\mathcal{B}\mathsf{Pol})^n(\mathcal{B}_0)$ and $\mathcal{B}_{n+1/2} = \mathsf{Pol}(\mathcal{B}_n)$.

Let $\mathcal{V}_0$ be the trivial class of languages $\{\emptyset, A^*\}$.
Let $\mathcal{V}_n = (\mathcal{B}\mathsf{Pol})^n(\mathcal{V}_0)$ and $\mathcal{V}_{n+1/2} = \mathsf{Pol}(\mathcal{V}_n)$.

Key result. These hierarchies are infinite [Brzozowski-Knast 1978].

Problem: Given n and a regular language L, decide whether L belongs to $\mathcal{B}_n$ (resp. $\mathcal{V}_n$, $\mathcal{B}_{n+1/2}$, $\mathcal{V}_{n+1/2}$).

Early results

- $\mathcal{V}_1$ is decidable (Simon 1972)
- $\mathcal{B}_1$ is decidable (Knast 1983)
- $\mathcal{V}_n$ is decidable iff $\mathcal{B}_n$ is decidable (Straubing 1985)
- $\mathcal{V}_{3/2} = \mathrm{Pol}\ \mathcal{V}_1$ is decidable (Arfi 1987, Pin-Weil 1995)

Connection with logic

The sentence $\exists i \ \mathbf{a}i$ defines the language A^*aA^*.

The sentence $\exists i \ \exists j \ \big((i < j) \wedge \mathbf{a}i \wedge \mathbf{b}j\big)$ defines the language $A^*aA^*bA^*$

$j = i + 1$ is a macro for
$(i < j) \wedge \forall k \ \Big((i < k) \rightarrow ((j = k) \vee (j < k))\Big)$.

$j \leqslant i$ is a macro for $j < i \vee j = i$.

The sentence $\exists j \ \forall i \ j \leqslant i \wedge \mathbf{a}j$ defines aA^*.

The sentence $\exists i \ \exists j \ j = i + 1 \wedge \mathbf{a}i \wedge \mathbf{a}j$ defines A^*aaA^*.

The hierarchies Σ_n, Π_n and Δ_n

Σ_n: Formulas $\exists^* \ \forall^* \ \exists^* \ \cdots \varphi$ with n alternating blocks of quantifiers.

Π_n: Formulas $\forall^* \ \exists^* \ \forall^* \cdots \varphi$ with n alternating blocks of quantifiers.

Δ_n: Formulas which are equivalent to a Σ_n-formula and to a Π_n-formula.

$\mathcal{B}\Sigma_n$: Boolean combinations of Σ_n-formulas.

Problem. Which languages are captured by these formulas?

Logical classes

Theorem (Thomas 1982, Perrin-Pin 1986)

(1) *The class* $\mathcal{B}\Sigma_n$ *captures* $\mathcal{V}_n$.

(2) *The class* Σ_n *captures* $\mathcal{V}_{n-1/2}$.

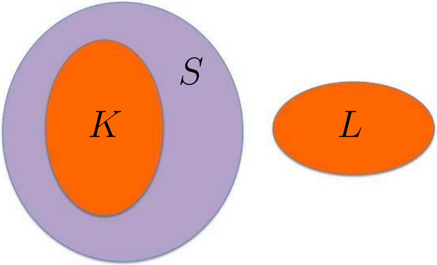
Theorem (Pin-Weil 1997)

The class Δ_n *captures* $\mathit{UPol}(\mathcal{V}_n)$.

What about decidability ?

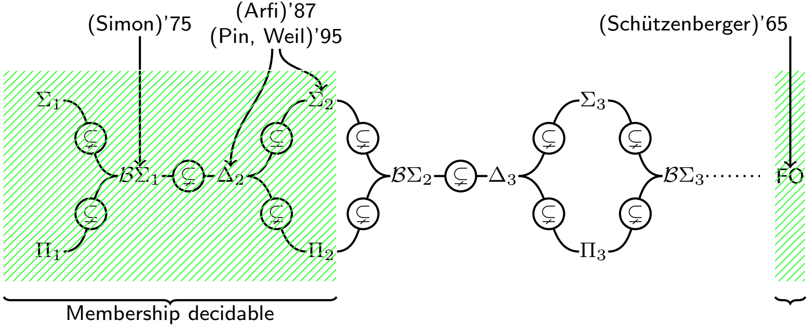
The separation problem

Let $\mathcal{C}$ be a class of regular languages. Is the following problem decidable: given two disjoint regular languages K and L, is there a language $S \in \mathcal{C}$ which separates K and L, that is, $K \subseteq S$ and $S \cap L = \emptyset$.

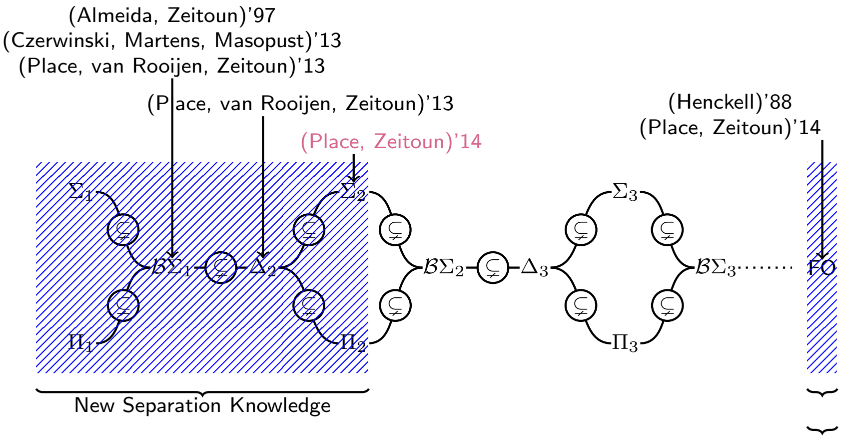


First order hierarchy

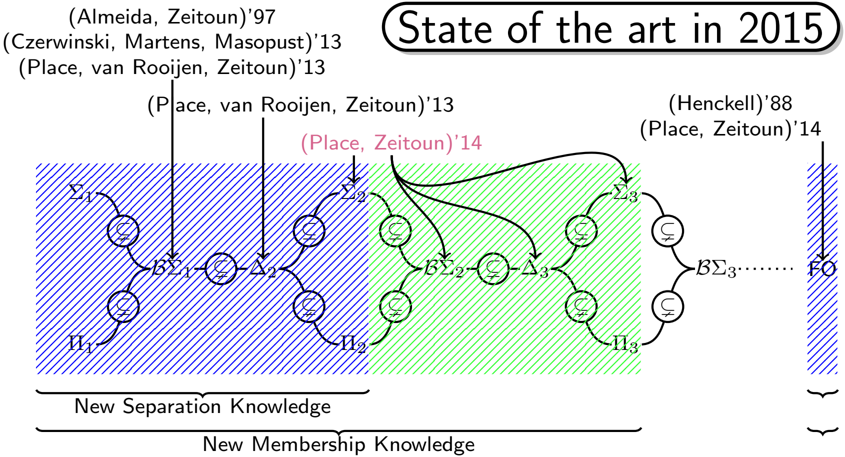
State of the art in 2013



First order hierarchy



First order hierarchy



Place, LICS'15: Separation for Σ_3 (hard), decidability for $\Delta_4, \Sigma_4, \Pi_4$

Still open for $\mathcal{B}\Sigma_3$

Almeida, Bartonova, Klíma, Kunc, DLT'15: Σ_n decidable implies Δ_{n+1} decidable.

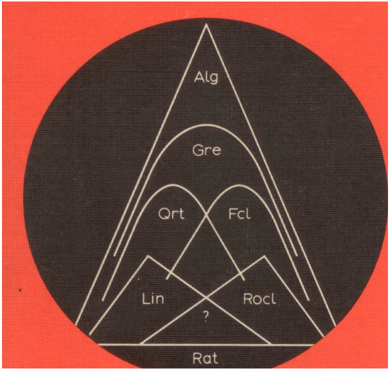
Conclusion: Janusz has excellent taste!

A striking selection of problems!

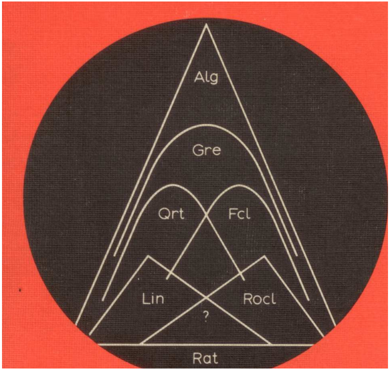
(1) Limitedness problem ✓
(2) Star height ??
(3) Restricted star height ✓
(4) Group complexity ?
(5) Star removal
(6) Regularity of non-counting classes $\sim$
(7) Optimality of prefix codes $\sim$
(8) Dot-depth hierarchy ↗

An early workshop on the dot-depth hierarchy

Janusz's opinion on cones of context-free languages



Janusz's opinion on cones of context-free languages



This picture is upside down! Rat should be at the top!