J.E. Pin (Paris)

Variétés de langages et combinatoire

Abstract: Let A be a finite alphabet.

The well-known Kleene's theorem states that a language L of A^* is rational iff its syntactic monoid is finite. Schützenberger's theorem states that a language L is star-free iff its syntactic monoid is group-free. It turns out that many subfamilies of the rational languages can be characterized in this way by properties of their syntactic monoids or semigroups. This lecture gives a survey of the various hierarchies of star-free languages, their descriptions in terms of semigroups, and the related decidability results and problems.

Let S,T denote semigroups

<u>Definition</u>: S *divides* T , $S < T$, if S is a quotient of a subsemigroup of T .

<u>Fact</u>: The divisibility relation $<$ is transitive ; moreover $<$ is an order on <u>finite</u> semigroups.

<u>Definition</u>: A *variety* of finite semigroups (monoids) is a class of finite semigroups (monoids) closed under taking subsemigroups, quotients and <u>finite</u> direct products, or equivalently: closed under division and finite direct products.

Let A denote a (finite) alphabet, then:
A^+ := the free semigroup over A ,
A^* := the free monoid over A .
A *language* (over A) is a subset $L \subseteq A^+$

<u>Definition</u>: A language $L \subseteq A^+$ is *recognized* by a semigroup S if there exists a semigroup morphism $\eta : A^+ \to S$ and a subset $P \subseteq S$ such that $L = P\eta^{-1}$.

<u>Example</u>: Let $\mathfrak{A}$ be an automaton recognizing L (in the automata-theoretic sense). Then the transition semigroup of $\mathfrak{A}$ recognizes L .

<u>Definition</u>: A language is *recognizable* if it is recognized by some finite semigroup.

<u>Fact</u>: Recognizable languages are closed under boolean operation ($\cap, \cup, \setminus$) and under inverse morphisms.

<u>Definition</u>: The *syntactic semigroup* $S(L)$ (syntactic monoid $M(L)$ resp.) of a language $L \subseteq A^+$ (A^* resp.) is the quotient of A^+ (A^* resp.) by the congruence $\sim_L$:

for $u,v \in A^+$ (A^* resp.)

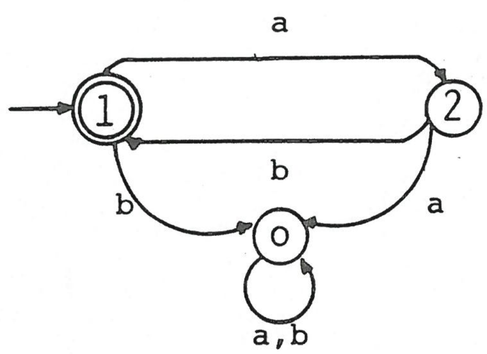
$u \sim_L v \iff (\forall x,y \in A^* : xuy \in L \iff xvy \in L)$

<u>Fact</u> 1) $S(L)$ recognizes L .

2) S recognizes L iff $S(L) < S$

<u>Construction</u>: $S(L)$ is the transition semigroup of the <u>minimal</u> automaton recognizing L .

<u>Example</u>: $A = \{a,b\}$, $L = (ab)^+ = \{ab, (ab)^2, (ab)^3, \ldots\}$



	1	2	0
a	2	0	0
b	0	1	0
a^2	0	0	0
ab	1	0	0
ba	0	2	0

in $S(L)$: $a^2 = b^2$, $aba = a$, $bab = b$, $(ab)^2 = ab$, $(ba)^2 = ba, \ldots$

<u>General Idea</u>: *Classify recognizable languages L by properties of $S(L)$ ($M(L)$ resp.)*

<u>Fact</u>: Let $\underline{V}$ denote a variety of semigroups (monoids resp.). The set of all languages of A^+ (A^* resp.) such that $S(L) \in \underline{V}$ ($M(L) \in \underline{V}$ resp.) is a boolean algebra $A^+\mathcal{V}$ ($A^*\mathcal{V}$ resp.) . $\mathcal{V} := \{A^+\mathcal{V} \mid A$ any finite alphabet$\}$ ($\mathcal{V} := \{A^*\mathcal{V} \mid A$ any finite alphabet$\}$ resp.)

$\mathcal{V}$ is the variety of languages associated to $\underline{V}$

EILENBERG'S THEOREM: Varieties of languages and varieties of semigroups (monoids resp.) are in 1-1 correspondence.

Rational languages of A :

1) $\emptyset$, $\{\Lambda\}$, $\{a\}$ where $a \in A$ are rational.

2) If L_1, L_2 are rational, $L_1 \cup L_2$, $L_1 \cdot L_2$ are rational.

3) If L is rational, L^* is rational.

KEENE'S THEOREM (1954): L is rational iff L is recognizable.

Star-free languages of A^* :

1) $\{a\}$ is star-free for all $a \in A$.

2) If L_1, L_2 are star-free, $L_1 \cdot L_2$ and each boolean combination of L_1, L_2 (including $A^* \setminus L_1$) are star-free.

Example: Let $A = \{a,b\}$.

$L = (a\, A^*\, b) \setminus (b\, b^*\, a)$ is star-free.

In fact: $\emptyset = \{a\} \setminus \{a\}$ is star-free

$A^* = A^* \setminus \emptyset$ is star-free

$b^* = A^* \setminus (A^* a\, A^*)$ is star-free

=> L is star-free .

SCHÜTZENBERGER'S THEOREM (1965): L is star-free iff M(L) is aperiodic, that is $\exists n \forall x \in M(L)\ x^n = x^{n+1}$ or equivalently:
If a group G divides M(L), then G is trivial.

Example: $(ab)^* = aA^* \cap A^*b \setminus (A^*aaA^* \cup A^*bbA^*)$,
where $A = \{a,b\}$.

Theorem: (Restivo 1973)

Let X be a finite code . X^* is star-free iff X^* is pure (i.e. $u^n \in X^* \Rightarrow u \in X^*$) .

Theorem (1979): For each aperiodic Monoid M there exists a finite pure code X such that $M < M(X^*)$

Fact: The set $\underline{A}$ of all finite aperiodic monoids is a variety.

Corollary: $\underline{A}$ is generated by the set

$\{M(X^*) \mid X \text{ is a finite pure code}\}$.

Definition: Let $u,v \in A^*$. u is a subword of v

if $v = v_o u_1 v_1 u_2 v_2, \ldots, u_n v_n$

$u = u_1, \ldots, u_n$ for some $n \geq 1$.

Example: aac and abcb are subwords of abaaacab .

Definition: $u,v \in A^*$. $u \sim_n v$ iff u and v

have the same set of subwords of length $\leq n$.

Example: anders $\sim_1$ andreas

ababab $\sim_3$ bababa

SIMON'S THEOREM (1972-1975): The following conditions are equivalent:

1) L is a union of $\sim_n$ -classes for some $n \geq o$ (i.e. piecewise testable)

2) L is in the boolean algebra generated by languages of the form $A^* a_1 A^* a_2 \ldots\ldots A^* a_n A^*$, $a_1, \ldots, a_n \in A$

3) $M = M(L)$ is J-trivial, that is, one of the following equivalent statements is true:

(a) $\forall a,b \in M \quad MaM = MbM \Rightarrow a = b$

(b) $\exists n \geq 0 \; \forall x,y \in M \; (xy)^n = (yx)^n$ and $x^n = x^{n+1}$.

Corollary (Straubing 1980): A monoid is $\mathcal{J}$-trivial iff it divides the monoid T_n of all boolean $n \times n$ matrices of the form

$$\begin{pmatrix} 1 & \varepsilon_{1,2} & \cdots & \varepsilon_{1,n} \\ & 1 & \ddots & \vdots \\ & & \ddots & \varepsilon_{n-1,n} \\ 0 & & & 1 \end{pmatrix}$$

for some $n > 0$.

Definition: $u, v \in A^*$. $u \simeq_n v$ if

u and v have the same prefix of length $< n$

and u and v have the same suffix of length $< n$.

Example: COMICS $\simeq_4$ COMBINATORICS

Theorem (Perrin 1971): Let $L \subseteq A^+$.

The following conditions are equivalent:

1) L is a union of $\simeq_n$ - classes for some $n > 0$ (i.e. endwise testable)

2) $L = XA^*Y \cup Z$ for some finite sets X, Y, Z A^+.

3) $S(L)$ is locally trivial (i.e. for all $e = e^2 \in S$ $eSe = \{e\}$).

Definition: For $u \in A^*$, $n > 0$ let

$$I_n(n) = \{\text{segments of } u \text{ of length } n\}$$

$u \equiv_n v$ iff $u \simeq_n v$ and $I_n(u) = I_n(v)$

Example: $u = ab\ aabaa\ ba \equiv_3 ab\ aa\ ba = v$

$I_3(u) = \{aba, baa, aab\} = I_3(v)$.

Theorem: (Brozowski-Simon 1973, McNaughton 1974)

The following conditions on $L \subseteq A^+$ are equivalent:

1) L is a union of $\equiv_n$-classes for some $n > o$ (i.e. locally testable).

2) L is in the boolean algebra generated by languages of the form uA^* , A^*v , A^*wA^* $(u,v,w \in A^+)$.

3) S(L) is locally a semilattice (that is: for all $e = e^2 \in S$ eSe is an idempotent and commutative monoid.).

<u>Theorem</u> (Restivo 1974): Let X be a finite code.
Then X is circular iff X^+ is locally testable.

<u>Fact</u>: The set of all semigroups, which are locally a semilattice, are a variety $\underline{L}\ \underline{J}_1$

<u>Theorem</u> (1979): For each $S \in \underline{L}\ \underline{J}_1$ there exists a finite circular code X such that $S < S(X^+)$.

<u>Corollary</u>: $\underline{L}\ \underline{J}_1$ is generated by the set

$\{S(X^+) \mid X \text{ is a finite circular code}\}$.

<u>Concatenation Hierarchies</u>.

F_o := "basic" boolean algebra

F_{n+1} := boolean algebra generated by languages of the form $L_o\, a_1\, L_1\, a_2 \ldots a_k\, L_k$ where $k \geq o$, $L_o,\ldots,L_k \in F_n$, $a_1,\ldots,a_k \in A$.

<u>Two main cases</u>:

1) Straubing's hierarchy (1981): $F_o = \{A^*, \emptyset\}$

2) Brzozowski-Cohen hierarchy or "dot-depth hierarchy" (1971) modified by Thêrien (1982)

$F_o = \{XA^*Y \cup Z \mid X,Y,Z \text{ finite} \subseteq A^+\}$.

<u>Theorem</u> (Straubing, Pin 1981): The following conditions are equivalent:

1) $L \in V_2$ (= second level of Straubing's hierarchy).

2) L is in the boolean algebra generated by languages of the form

$$A_o^* a_1 A_1^* \dots a_k A_k^* \quad , k \geq o$$
$$A_i \subseteq A \ , \ a_i \in A \ .$$

3) $M(L) \in \underline{PJ}$, the variety of monoids generated by all power monoids of the form

$P(M)$, $M \in \underline{J} := \{M \mid M \text{ is } J\text{-trivial}\}$.

<u>Corollary</u>: $M \in \underline{PJ}$ iff M divides the monoid K_n of all $n \times n$ boolean upper-triangular matrices for some $n > o$.

<u>OPEN QUESTION</u>: Is $\underline{PJ}$ decicable?

That is, does there exist an algorithm to test membership in $\underline{PJ}$?

<u>Knast condition</u> (k):

For all idempotents $e_1, e_2 \in S$, for all $x, y, u, v \in S$

$$(e_1xe_2y)^n \, e_1xe_2ve_1(ue_2ve_1)^n = (e_1xe_2y)^n e_1(ue_2ve_1)^n$$

<u>Theorem</u> (Knast, to appear):

$L \in B_1$ (L has dot-depth ≤ 1) iff

S(L) satisfies (K) .

<u>Fact</u>: $P(S_1^1 x \dots x S_n^1)$ is a semiring.

($S_1, \dots, S_n$ are semigroups) .

Definition: The Schützenberger product

$\Diamond_n(S_1,\ldots,S_n)$ of semigroups $S_1,\ldots,S_k$ is the set of upper-triangular matrices $(M_{i,j})$ with entries in $\mathcal{P}(S_1^1 x,\ldots,x S_n^1)$, such that :

1) $M_{ii} = \{(\Lambda,\ldots,\Lambda,s_i,\Lambda,\ldots,\Lambda)\}$ (s_i being the i-th-component)

for some $s_i \in S_i$

2) $M_{ij} \subseteq \{(s_1,\ldots,s_n) \in S_1^1 x \ldots x S_n^1 \mid s_1 = \ldots = s_{i-1} = 1 = s_{j+1} = \ldots = s_n\}$

WARNING !

$$\Diamond_2(\Diamond_2(S_1,S_2),S_3) \neq \Diamond_3(S_1,S_2,S_3)$$

$$\Diamond_2(\Diamond_2(S_1,S_2),S_3) \neq \Diamond_2(S_1,\Diamond_2(S_2,S_3))$$

$$\Diamond_2(S_1,\Diamond_2(S_2,S_3)) \neq \Diamond_3(S_1,S_2,S_3)$$

Fact (Straubing 1981): If $L_o,\ldots,L_n$ are recognized by $S_o,\ldots,S_n$ respectively and if $a_1,\ldots,a_n$ are letters, then $L_o\ a_1\ L_1\ \ldots\ a_n\ L_n$ is recognized by $\Diamond_{n+1}\ (S_o,\ldots,S_n)$.

Theorem (Reutenauer $n=1$ 1979, Pin 1981).

If $L \subseteq A^+$ is recognized by $\Diamond_{n+1}(S_o,\ldots,S_n)$, then L is in the boolean algebra generated by languages of the form $L_{i_o}\ a_i\ L_{i_1},\ldots,a_r\ L_{i_r}$,

$o \leq i_o < i_1 < \ldots < i_r \leq n$, a_k letters and L_{i_k} recognized by S_{i_k} $(k = 1,\ldots,r)$.

Definition: Let $\underline{V}$ be a variety . $\Diamond(\underline{V})$ is the variety generated by all semigroups $\Diamond_n(S_1,\ldots,S_n)$ for $n > o$, $S_i \in \underline{V}$.

Theorem (Brzozowski-Knast 1978. Straubing 1981):

Brzozowski's hierarchy is infinite.

The facts on Straubing's and Brzozowski's hierarchies are summarized in the following tables.

STRAUBING'S HIERARCHY

Level	Languages $\mathcal{V}_n$	Varieties of monoids $\underline{V}_n$
0	$\emptyset, A^*$	$\underline{V}_o = \underline{I}$ $(=\{\{1\}\})$ trivial variety.
1	Boolean algebra generated by languages of the form $A^* a_1 A^* \ldots\ldots a_n A^*$	$\underline{V}_1 = \underline{J}$ (J-trivial)
2	Boolean algebra generated by languages of the form $A_o^* a_1 A_1^* a_2 \ldots a_n A_n^*$, $A_i \subseteq A$, $a_i \in A$	$\underline{V}_2 = \underline{PJ}$ (Power monoids of monoids in $\underline{J}$)
3 ↓ . . . . . . infinite hierarchy		$\underline{V}_{n+1} = \Diamond(\underline{V}_n)$

BRZOZOWSKI'S HIERARCHY

Level	Languages $\mathcal{B}_n$	Varieties of semigroups $\underline{B}_n$
0	$XA^*Y \cup Z$, X,Y,Z finite subsets of A^+	$\underline{B}_o = \underline{LI}$ (= locally trivial semigroups. (eSe = e)).
1	Boolean algebra generated by language of the form $w_o A^* w_1 A^* , \ldots , w_n A^* w_{n+1}$ $w_i \in A^+$	$\underline{B}_1$ = semigroups satisfying (K)
2 ↓ . . . . infinite hierarchy		$\underline{B}_{n+1} = \Diamond(\underline{B}_n)$

Definition: Let T be a semigroup (written multiplicatively) and S be a semigroup (written additively, without being commutative in general) and $T \times S \longrightarrow S$

$$(t,s) \longmapsto t \cdot s$$

satisfying

1) $t \cdot (s_1+s_2) = t \cdot s_1 + t \cdot s_2$
2) $(t_1, t_2) \quad s = t_1(t_2 s)$.

The semidirect product $S * T$ is defined on $S \times T$ by

$$(s,t)(s',t') = (s+t\,s', t\,t') .$$

For two varieties $\underline{V}$, $\underline{W}$:

$\underline{V} * \underline{W}$ = variety generated by all semidirect products $S * T$, $S \in \underline{V}$, $T \in \underline{W}$.

Theorem (Straubing 1982):

$\underline{B}_n = \underline{V}_n * \underline{LI}$
for all $n \geq o$.

Theorem (Margolis-Straubing, 1982):

$\underline{B}_n$ is decidable iff $\underline{V}_n$ is decidable.

Bibliographie:

J. A. Brzozowski: Hierarchies of aperiodic languages, RAIRO, Informatique Théorique, vol. 10, 1976, 33-49

J. A. Brzozowski et R. Knast: The dot-depth hierarchy of star-free languages is infinite, J. Computer and System Sciences, vol. 16, 1978, 37-55

J.A. Brzozowski et I. Simon: Characterizations of locally testable events, Discrete Mathematics, vol. 4, 1973, 243-271.

S. Eilenberg: Automata, languages and machines, vol. B, Academic Press, New York (1976)

R. Knast: Some theorems on graph congruences. A paraître dans la RAIRO, Informatique Théoretiqe

R. Knast: A semigroup characterization of dot-depth one languages. A paraître dans la RAIRO, Informatique Théorique.

G. Lallement: Semigroups and Combinatorial applications, Wiley, N.Y., 1979

J.E. Pin: Variétés de langages et variétés de semigroups, Thése, Paris, 1981

J.E. Pin et J. Sakarovitch: Une application de la représentation matricielle des transductions. A paraître.

J.E. Pin et H. Straubing: Monoids of upper-triangular matrices, à paraître.

C. Reutenauer: Sur les variétés de langages et de monoïdes, Lect. Notes in Computer Sc. no. 67, Springer Berlin (1979) 260-265

I. Simon: Hierarchies of events with dop-depth one, These, Université de Waterloo (1972)

I. Simon: Piecewise testable events, Lect. Notes in Computer Science no. 33, Springer Berlin (1975), 214-222

H. Straubing: A generalization of the Schützenberger product of finite monoids, Theor. Comp. Sc. 13 (1981), 137-150

H. Straubing: A study of the dot-depth hierarchy (à paraître).