UNIVERSITY OF WATERLOO
COMPUTER SCIENCE DEPARTMENT

Open Problems About Regular Languages

Janusz Brzozowski

CS-80-03

January, 1980

UNIVERSITY OF WATERLOO
COMPUTER SCIENCE DEPARTMENT

Open Problems About Regular Languages

Janusz Brzozowski

Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1

ABSTRACT

We discuss six relatively old open problems about regular languages, namely:

Star Height
Restricted Star Height
Group Complexity
Star Removal
Regularity of Noncounting Classes
Optimality of Prefix Codes

Presented at the International Symposium on Formal Language Theory, Santa Barbara, California, December 10-14, 1979.

Open Problems About Regular Languages

Janusz Brzozowski

Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1

0. Introduction

The theory of regular languages and finite automata was developed in the early 1950's, and is therefore one of the oldest branches of theoretical computer science. Regular languages constitute the best known family of formal languages, and finite automata constitute the best known family of abstract machine models. The concepts of regular languages and finite automata appear very frequently in theoretical computer science, and have several important applications. There is a vast literature on these subjects.

In spite of the fact that many researchers have worked in this field there remain several difficult open problems. Six of these problems are discussed in this paper. There are more than six open problems about regular languages; the choice of these six represents the personal prejudices of the author. It is not our intention here to imply that other open problems are not significant. However, the problems chosen do appear to be of fundamental importance and considerable difficulty. Most of them are intimately involved with the fundamental property of finite automata, namely finiteness.

For the most part we have adopted the terminology and notation of Eilenberg [13,14].

1. Star Height

In a monograph [21] published in 1971 McNaughton and Papert include a collection of open problems concerning regular languages. Their list is headed by the star height problem. To illustrate their uncertainty about the problem, I quote their final paragraph:

Research supported in part by the Natural Sciences and Engineering Research Council of Canada under Grant No. A-1617. Preparation of this paper was supported in part by the National Science Foundation of the United States under Grant No. MCS79-04012.

> At this moment we are unwilling to conjecture even that there exist events of general loop complexity two or more. The entire question of events whose general loop complexity exceeds one is wide open. Only one of our conjectures remains credible: that if there exist regular events of loop complexity n, for $n \geqslant 2$, then there exist regular events of loop complexity n whose syntactic monoids are groups. We suspect that someone might prove this hypothetical statement without answering the main question.

After eight years not much has changed. I find it surprising that no progress had been made on such an intriguing question. The interest in the theory of finite automata and regular languages as a research topic for computer scientists has decreased significantly over the last 15 years. This is understandable since the problems that are left are quite difficult and perhaps unfair as Ph.D. thesis topics. However, the number of people that were actively involved in the area was very large indeed, and I would expect that a handful of them would retain an active interest in this problem, as a hobby. Yet there appear to be no new results.

Let Σ be a finite alphabet and Σ^* the free monoid generated by Σ with unit element 1. *Regular expressions* over Σ are defined inductively:

(a) ϕ, 1 and σ for each $\sigma \in \Sigma$ are regular expressions.

(b) If E and F are regular expressions, then so are

$$\overline{E},\ E \cup F,\ E \cap F,\ EF \text{ and } E^*.$$

In this definition ϕ denotes the empty language and $\overline{E} = \Sigma^* - E$. The remaining operators are union, intersection, concatenation and star.

The *(star) height* Eh of a regular expression E is defined inductively as

(a) $\phi h = 1h = 0$, and $\sigma h = 0$ for all $\sigma \in \Sigma$.

(b) If E and F are regular expressions then

$$\overline{E}h = Eh,$$

$$(E \cup F)h = (E \cap F)h = (EF)h = \max\{Eh, Fh\},$$

$$E^*h = 1 + Eh.$$

In other words Eh is the maximum number of nested stars in E. For example, if $E = (\sigma \cup \tau\tau^*\sigma)^*$, then $Eh = 2$.

If $A \subset \Sigma^*$ is a regular language, the height Ah of A is the least height of a regular expression denoting A. If E is a regular expression let $|E|$ be the language denoted by E. In the example above $|E|$ is of height zero because

$$|E| = |1 \cup \overline{\phi}\sigma|$$

and $(1 \cup \overline{\phi}\sigma)h = 0$.

The family of languages of height zero was characterized in 1965 by Schützenberger [26] who showed that $Ah = 0$ iff the syntactic monoid M of A is

aperiodic, i.e. has only trivial subgroups. This family of star-free languages is relatively well-known. Apart from this, we know that if M contains a non-trivial group then $Ah > 0$, but it is not known whether there are any languages of height two!

As McNaughton and Papert reported [21], for many years before 1971 the language $|\mathcal{A}_1|$ accepted by the automaton $\mathcal{A}_1$ of Figure 1 was thought to be of height two.

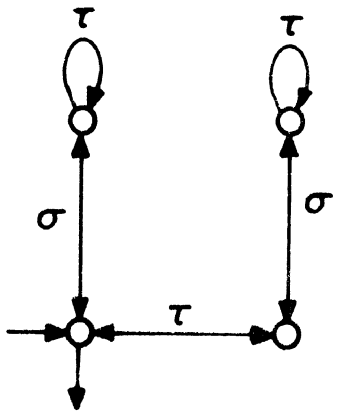

Fig. 1. Automaton $\mathcal{A}_1$

Henneman [18] showed that it is of height one. In fact an expression of height one for $A_1 = |\mathcal{A}_1|$ is

$$A_1 = [E \cap F] \cup [(\tau \cup \sigma\tau^*\sigma)E \cap (\tau^*\sigma)^2 F],$$

where $E = ((\tau \cup \sigma\tau^*\sigma)^2)^*$, and $F = (\tau \cup (\tau^*\sigma)^4)^*$. A natural height-two expression for A_1 is:

$$(\sigma\tau^*\sigma \cup \tau(\sigma\tau^*\sigma)^*\tau)^*,$$

where $\tau^* = \overline{(\overline{\phi}\ \sigma\ \overline{\phi})}$. Informally, one can view $\mathcal{A}_1$ as counting σ's modulo 2 and counting τ's modulo 2, but only those τ's that occur after an even number of σ's. A more complicated example suggested by Thérien [29] is

$$A_2 = (\sigma\tau^*\sigma \cup \tau\sigma^*\tau(\sigma\tau^*\sigma)^*\tau\sigma^*\tau)^*,$$

accepted by the six-state automaton of Figure 2. This language is suspected of being a height-two language.

The language A_1 of Figure 1 is very closely related to A_3 accepted by $\mathcal{A}_3$ of Figure 3. This language has a very natural description as follows.

Let $u, w \in \Sigma^*$ with $u = \sigma_1 \ldots \sigma_n$, where $\sigma_1, \ldots, \sigma_n \in \Sigma$. The *binomial coefficient* $\binom{w}{u}$ is the number of factorizations

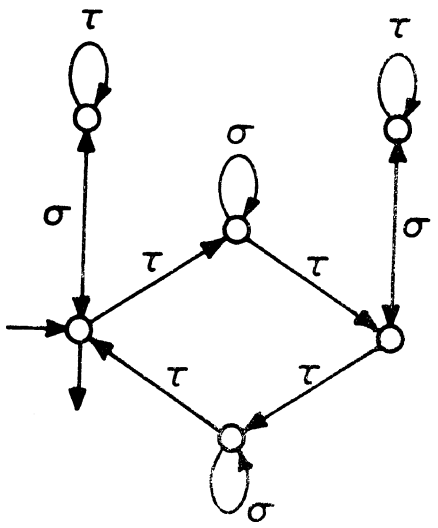

Fig. 2. A_2

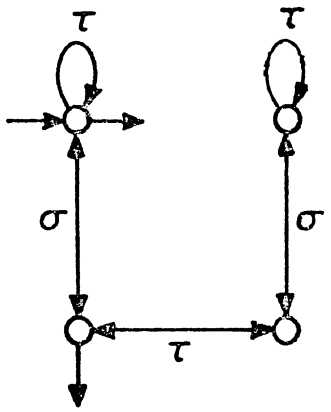

Fig. 3. A_3

$$w = v_0\sigma_1 v_1\sigma_2 \ldots v_{n-1}\sigma_n v_n$$

with $v_0, \ldots, v_n \in \Sigma^*$. One can verify that

$$A_3 = \{w \mid w \in \{\sigma, \tau\}^* \text{ and } \binom{w}{\sigma\tau} \equiv 0 \bmod 2\}.$$

In other words, A_3 counts modulo 2 the number of ways in which $\sigma\tau$ is a subword of w. A height-one expression is easily obtainable for A_3 from A_1. In fact

$$A_3 = \tau^* \cup \tau^*\sigma A_1(1 \cup \sigma\tau^*).$$

A natural generalization of A_3 is

$$A_4 = \{w \mid w \in \{\sigma, \tau, \eta\}^* \text{ and } \binom{w}{\sigma\ \tau\ \eta} \equiv 0 \bmod 2\}.$$

The reduced automaton accepting A_4 is shown in Figure 4. At the present time A_4 is another candidate for height two.

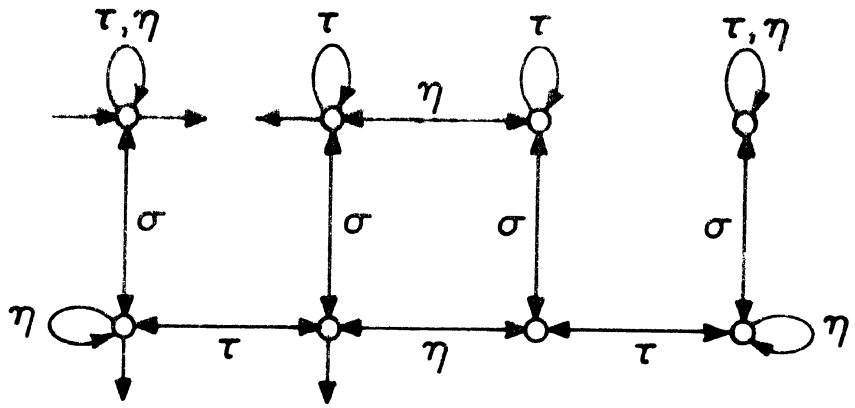

Fig. 4. A_4

To the best of our knowledge, the only published work on the star height problem (other than Schützenberger's paper on star-free languages) is that of Henneman [18]. His main results are summarized below. Let G be the syntactic monoid of the language A, and let the order of G be n. Then

if G is a group,	$Ah \leqslant n$
if G is a solvable group,	$Ah \leqslant \log n$
if G is a supersolvable group,	$Ah \leqslant 2 + \log\log n$
if G is an abelian group,	$Ah \leqslant 1$

Henneman defines the height $\mathcal{A}h$ of a complete (deterministic) automaton $\mathcal{A}=(Q,i,T)$ to be the height of the language $|\mathcal{A}|$ accepted by $\mathcal{A}$. The height Sh of a semiautomaton S (automaton without initial state and final states) is the maximum height of any automaton associated with S. Let $S(q,q') = \{w \mid w \in \Sigma^* \text{ and } qw = q'\}$ in the semiautomaton S. One easily verifies that the height of S is equal to the maximum height of the languages $S(q,q')$ over all $q,q' \in Q$.

Given a semiautomaton S with state set Q and input alphabet Σ, let M be the transformation monoid associated with S. Then the monoid semiautomaton of S is a semiautomaton S_M with state set M and input alphabet Σ.

Henneman shows that $Sh = S_M h$. To illustrate an application of this result consider Figure 5. One verifies that S_M is isomorphic to the monoid semiautomata of both S_1 and S_2. Hence $S_1 h = S_M h = S_2 h$. It is easy to verify that $S_2 h = 1$; we can then conclude also that $S_1 h = 1$.

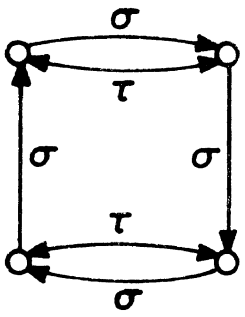

Fig. 5(a). S_1

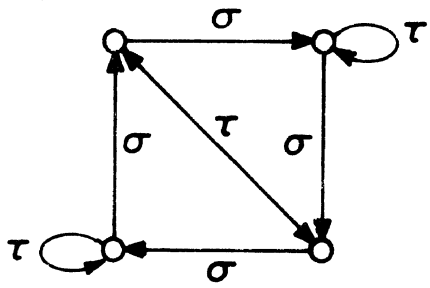

Fig. 5(b). S_2

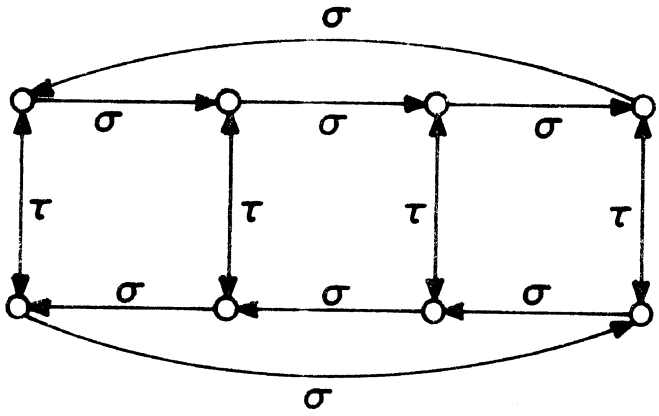

Fig. 5(c). S_M

The bounds on star height that were mentioned above apply only to languages whose syntactic monoids are groups. In that case the corresponding semiautomata are permutation semiautomata. Henneman observes that, for a permutation semiautomaton S, $Sh = (S(q,q))h$ where q is any state of S. Hence one needs to consider only one language associated with any semiautomaton.

In the general case, if it is known only that the syntactic monoid of a language A is a group G of order n, then the monoid automaton has n states. It is easy to find expressions of height $\leqslant n$ for any n-state automaton, using for example the McNaughton-Yamada algorithm [22]. If the reduced automaton for

A has m states and $m < n$, then of course the bound can be lowered to m.

In case G is abelian, it is not difficult to write a height-one expression which is a finite union of expressions of the form

$$\bigcap_{\sigma \in \Sigma} A_\sigma^{p_\sigma} (A_\sigma^{q_\sigma})^*,$$

where $A_\sigma = \overline{(\overline{\phi}\ \sigma\ \overline{\phi})}\ \sigma\ \overline{(\overline{\phi}\ \sigma\ \overline{\phi})}$ is the set of all words over Σ containing exactly one σ, and $p_\sigma, q_\sigma \geqslant 0$.

The results concerning solvable and supersolvable groups are deduced from some classical results from group theory and the following theorem of Henneman:

Theorem: Let G be a group and N a normal abelian subgroup of G. Then

$$Gh \leqslant (G/N)h + 1.$$

(The height of a group G is the height of the monoid semiautomaton with state set G, input alphabet G, and multiplication in G as the transition function.)

Henneman's thesis closes with some techniques for proving that certain two-input semiautomata are of star height one, and with a list of open problems.

The fundamental question remains: Is the star height hierarchy finite or infinite? Our ignorance about this problem is well illustrated by the fact that we are not even able to answer a much simpler question: Is there a language of height two?

Henneman's results lead one to believe that the star height of a language is a property that can be characterized in its syntactic monoid. Also, almost all known characterizations of subclasses of the class of regular languages can be done in the framework of Eilenberg's variety theory [14]. The following observation shows that this approach can work only if all languages are of height $\leqslant 1$. Let H_i be the class of all languages of height $\leqslant i$, for $i \geqslant 0$.

Proposition: H_1 is a $*$-variety iff every regular language is of height $\leqslant 1$.

Proof: Suppose $A \in H_i$ for $i > 1$, and let M_A be the syntactic monoid of A. By a recent theorem of Pin [24], there exists a finite language B such that M_A divides M_{B^*}. However, $B^* \in H_1$. If H_1 is a $*$-variety, it follows by Eilenberg's theorem that also $A \in H_1$. Hence all regular languages are of height $\leqslant 1$, if H_1 is a $*$-variety. The converse holds since the class of all regular languages is a $*$-variety.

Note that membership in H_0 *is* a property determined by the syntactic monoid. Also, H_0 *is* a $*$-variety.

2. Restricted Star Height

A *restricted* regular expression is a regular expression without intersections and complements. The *restricted star height* of a regular language A is the minimum height of a restricted regular expression E denoting A. Our knowledge of the restricted star height problem is considerably better than it is for the problem of the last section. It is known that the restricted height hierarchy is

infinite, and algorithms for finding the restricted height exist for several families of languages. However, the general case is still open; i.e., given an arbitrary regular language, it is not known how to find its restricted height. Throughout this section height means restricted height.

The restricted star height problem was introduced by Eggan [12] in 1963. He showed that for each $h \geqslant 0$ there exists a language A_h over alphabet Σ_h which has height h; the size of the alphabet Σ_h grows with h. Eggan raised the question whether there exist languages of arbitrary height over the two letter alphabet. (Languages over a one letter alphabet are all of height 0 or 1.) The question was answered positively by McNaughton (unpublished notes) and later by Dejean and Schützenberger [11] in 1966. Let $\mathbf{A}_h = (Q, i, \{i\})$ be the automaton over the alphabet $\Sigma = \{\sigma, \tau\}$, where $Q = \{0, 1, \ldots 2^h - 1\}$, $q\sigma = q + 1 \ mod \ 2^h$ and $q\tau = q - 1 \ mod \ 2^h$ for all $q \in Q$. Dejean and Schützenberger showed that $|\ \mathbf{A}_h\ |$ is of height h.

Eggan related star height of a language to the notion of *cycle rank* of a graph representing the language. The rank of a graph is a measure of the loop complexity of the graph; the precise definition is somewhat involved. Eggan showed that for every regular language A there exists a transition graph (a finite-automaton-like object which permits empty word transitions) whose rank is Ah. Cohen [6,9] showed that the search can be limited to nondeterministic automata, namely that for each regular language A there exists a nondeterministic automaton whose rank is Ah. We will briefly describe one of Cohen's approaches.

In what follows we need to have an explicit notation for the transitions in a finite automaton. Thus we will use the notation $\mathbf{A} = (Q, I, T, E)$ where $E \subset Q \times \Sigma \times Q$. For any set S, $\hat{S}$ denotes the set of all subsets of S.

Let $\underset{\sim}{\mathbf{A}} = (Q, i, T, E)$ be a minimal deterministic Σ-automaton. Let $\tilde{\mathbf{A}} = (\tilde{Q}, \tilde{I}, \tilde{T}, \tilde{E})$ be the nondeterministic *subset automaton* derived from $\mathbf{A}$ as follows:

$$\tilde{Q} = \hat{Q} - \phi$$
$$\tilde{I} = \{X \mid X \in \hat{Q},\ i \in X\}$$
$$\tilde{T} = \{X \mid X \in \hat{T} - \phi\}$$
$$\tilde{E} = \{(X, \sigma, X') \mid X\sigma \subset X'\}$$

where $X\sigma = \{q' \in Q \mid (q, \sigma, q') \in E \text{ for some } q \in X\}$, as usual. One verifies that, if $X \xrightarrow{s} X'$ is a path in $\tilde{\mathbf{A}}$ for some $s \in \Sigma^*$, then $Xs \subset X'$. It follows that $s \in |\ \tilde{\mathbf{A}}\ |$ implies $Xs \subset X'$ for some X such that $i \in X$ and $X' \subset T$. In particular, $is \in T$ or $s \in |\ \mathbf{A}\ |$. Hence $|\ \tilde{\mathbf{A}}\ | \subset |\ \mathbf{A}\ |$. It is easily seen that $|\ \mathbf{A}\ | \subset |\ \tilde{\mathbf{A}}\ |$; thus $|\ \tilde{\mathbf{A}}\ | = |\ \mathbf{A}\ |$.

Next let $k > 0$ and let $\tilde{\mathbf{A}}_k = (\tilde{Q} \times \mathbf{k}, \tilde{I} \times \mathbf{k}, \tilde{T} \times \mathbf{k}, \tilde{E}_k)$, where $\mathbf{k} = \{0, 1, \ldots, k-1\}$, be the nondeterministic *k-subset automaton* with

$$\tilde{E}_k = \{((X, j), \sigma, (X', j')) \mid (X, \sigma, X') \in \tilde{E}\}.$$

One easily verifies that $|\ \tilde{\mathbf{A}}_k\ | = |\ \mathbf{A}\ |$. Note that $\tilde{\mathbf{A}}_1$ is isomorphic to $\tilde{\mathbf{A}}$.

Let $\mathcal{A}_k = (Q_k, I_k, T_k, E_k)$ be a nondeterministic automaton,

where $Q_k \subset \tilde{Q} \times \mathbf{k}$, $I_k \subset \tilde{I} \times \mathbf{k}$, $T_k \subset \tilde{T} \times \mathbf{k}$ and $E_k \subset \tilde{E}_k$.

In other words, the graph of $\mathcal{A}_k$ is a subgraph of the graph of $\tilde{\mathcal{A}}_k$. We will say that any $\mathcal{A}_k$ satisfying the conditions above is a *k-graph* of $\mathcal{A}$. Clearly $|\mathcal{A}_k| \subset |\mathcal{A}|$. Without loss of generality we can assume that $\mathcal{A}_k$ is trim (i.e. every state appears in some successful path).

Cohen showed that any nondeterministic automaton recognizing a given language A can be viewed as a k-graph of $\mathcal{A}$, where $\mathcal{A}$ is the minimal deterministic automaton recognizing A. For let $\mathcal{B} = (P, I_B, T_B, E_B)$ be any trim nondeterministic automaton for A. Let

$$B_p = \{s \mid s \in \Sigma^*,\ p \in I_B s\}$$

for any $p \in P$. Define the function $f : P \to \hat{Q}$ by

$$pf = iB_p.$$

One can verify that

$$p \in P \text{ implies } pf \neq \phi;$$

$$p \in I_B \text{ implies } i \in pf;$$

$$p \in T_B \text{ implies } pf \subset T;$$

$$(p, \sigma, p') \in E_B \text{ implies } (pf)\,\sigma \subset p'f.$$

Thus all the conditions are satisfied for $\mathcal{B}$ to be isomorphic to a k-graph of $\mathcal{A}$.

As an example, consider the minimal deterministic automaton $\mathcal{A}$ of Figure 6(a) over $\Sigma = \{\sigma, \tau, \eta\}$, and the nondeterministic automaton $\mathcal{B}$ of Figure 6(b). The set of states of $\mathcal{A}$ associated with each node of $\mathcal{B}$ is shown in Figure 6(b). Note that $\{q\}$ appears twice; hence $\mathcal{B}$ is a 2-graph of $\mathcal{A}$. The cycle rank of $\mathcal{B}$ is 1 and, if A is the language of $\mathcal{A}$, we conclude that $Ah = 1$. One can verify that there does not exist any 1-graph of $\mathcal{A}$ that is of rank 1; i.e. it is necessary to use a 2-graph in order to display the height of A. In fact for each $k \geqslant 1$ an example can be produced where it is necessary to go to a k-graph. The question is whether k can be bounded. The following conjecture of Cohen is still open:

Conjecture: If $\mathcal{A}$ has m states then there exists a $(2^m - 1)$-graph of $\mathcal{A}$ whose rank is equal to the height of A.

The interested reader is referred to the literature [1,6-12,17-20] for further results about star height.

3. Group Complexity

In the first two problems we have considered two rather direct measures of complexity of a regular language. It is by now a well-established fact that many properties of languages are reflected in the properties of the corresponding semigroups. The problem discussed here deals with the complexity of semigroups.

A transformation semigroup $X = (Q, S)$, abbreviated *ts*, consists of a finite set Q and a subsemigroup S of $PF(Q)$, where $PF(Q)$ is the monoid of all partial

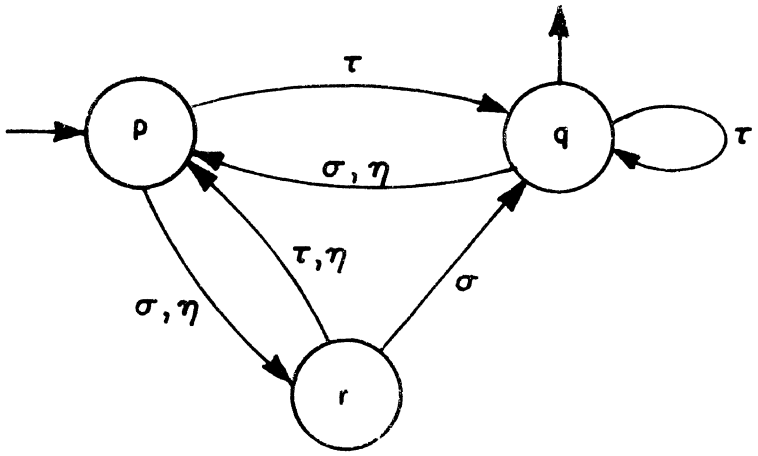

Fig. 6(a). A

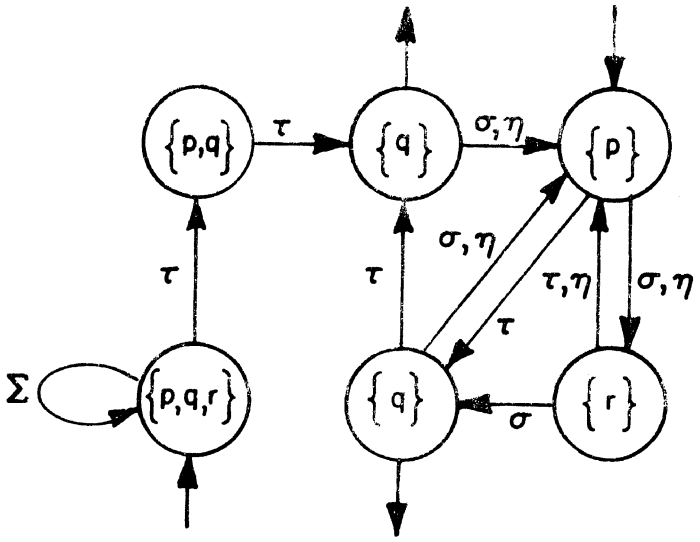

Fig. 6(b). B

functions $Q \to Q$ with composition as multiplication. If X and Y are *ts*'s, then $X < Y$ denotes the relation X *divides* Y and $X \circ Y$ is the wreath product of X and Y. It is well known that any *ts* $X = (Q, S)$ has a decomposition

$$X < A_{n+1} \circ G_n \circ A_n \circ \ldots \circ G_1 \circ A_1, \tag{1}$$

where A_i is an aperiodic (group-free) *ts* for $i = 1, \ldots, n+1$, and G_i is a group for $i = 1, \ldots, n$. The *(group) complexity* Xc of X is the smallest integer n over all decompositions of type (1). The Krohn-Rhodes Decomposition Theorem guarantees the existence of such a decomposition for each *ts* X. Moreover, G_i can be chosen to be a simple group such that $G_i < S$. However, in the definition of complexity, these conditions are not imposed; any finite group may be used in (1).

The open problem is: Does there exist an algorithm for finding the complexity of a given *ts*? Intuitively one would expect that, given the cardinality m of S, one should be able to eliminate those groups that are "too large" for S and limit the search to groups of cardinality $\leqslant$ some bound depending on m. However, no such bound has been found.

Notice that a *ts* has group complexity zero iff it is aperiodic. Every *ts* that is a nontrivial group has complexity one. The group complexity hierarchy is infinite because it can be shown that for $n \geqslant 1$ the *ts* $X_n = (\mathbf{n}, F_n)$, where F_n is the monoid of all functions $\mathbf{n} \to \mathbf{n}$, has complexity $n-1$.

The complexity problem was introduced by Krohn and Rhodes in 1965. A rather detailed account of the problem is given by Tilson in a chapter of Eilenberg's Vol. B [14].

An interesting connection between a related complexity measure and a certain hierarchy of languages has been recently made by Straubing [28]. He showed that any transformation monoid X containing only solvable groups has a decomposition

$$X < K_n \circ K_{n-1} \circ \ldots \circ K_1, \tag{2}$$

where for each i, K_i is either aperiodic or an abelian group. Let Xa be the *abelian group complexity* of X, i.e. the smallest number of abelian groups over all decompositions of type (2).

Let $\mathcal{V}_j$ be the family of all languages that can be constructed from the letters of the alphabet using boolean operations, concatenation, and $\leqslant j$ levels of a certain counting operation [28]. Straubing shows that a language A is in $\mathcal{V}_j$ iff its syntactic monoid has abelian complexity $\leqslant j$.

4. Star Removal

In 1965 Paz and Peleg [23] asked whether every regular language can be decomposed as a product of a finite number of stars and primes, where a subset A of Σ^* is prime iff $A = BC$ implies $B = 1$ or $C = 1$, and it is a star iff $A = A^*$. A positive answer to this question was given by Brzozowski and Cohen [2] in 1967. The decomposition procedure described in [2] does not lead to a unique factorization as shown by the following example:

$$\begin{aligned} A &= (1 \cup \sigma_0)[\sigma_1\sigma_0 \cup (\sigma_3 \cup \sigma_2\sigma_1)(\sigma_2\sigma_1)^*(\sigma_1\sigma_0 \cup \sigma_0)]^* \\ &\qquad (1 \cup \sigma_3 \cup \sigma_2\sigma_1)(\sigma_2\sigma_1)^*\sigma_1 \\ &= (1 \cup \sigma_3)[\sigma_2\sigma_1 \cup (\sigma_0 \cup \sigma_1\sigma_0)(\sigma_1\sigma_0)^*(\sigma_2\sigma_1 \cup \sigma_3)]^* \\ &\qquad (1 \cup \sigma_0 \cup \sigma_1\sigma_0)(\sigma_1\sigma_0)^*\sigma_1. \end{aligned}$$

In both cases all the factors are either stars or primes.

In contrast to this, the ideas of [2] and [23] suggest a natural procedure leading to a unique decomposition as described below.

Let $\mathcal{A} = (Q,i,T)$ be a deterministic Σ-automaton and let $A = |\mathcal{A}|$. Consider the Σ-automaton $\mathcal{B} = (Q,i,T')$ where the transition function of $\mathcal{B}$ is the same as that of $\mathcal{A}$ and

$$T' = \{q \mid qA \subset T\}.$$

We will show that $B = |\mathcal{B}|$ is the maximal left star factorable from A. Since $iA \subset T$, we have $i \in T'$ and $1 \in B$. If $s,t \in B$ then $is,it \in T'$ and $isA, itA \subset T$. The last condition is equivalent to $tA \subset A$. Thus $istA \subset isA \subset T$, $ist \in T'$ and $st \in B$. Altogether B is a star. It is also easy to check that $A = BA$. Now suppose $A = B'C$ for some star B'; then $A = B'B'C = B'A$. If $s \in B'$, then $sA \subset A$, $isA \subset T$, $is \in T'$ and $s \in B$. It follows that $B' \subset B$, i.e. that B is maximal.

One easily verifies that, given any decomposition $A = BD$ where B is a star, there is a unique minimal tail C of A with respect to B (such that $A = BC$), namely

$$C = A - (B-1)A.$$

The following procedure then suggests itself. Given a regular language A, find its maximal left star B_1 and its minimal tail C_1 with respect to B_1. Repeat, replacing A by C_1, etc. After n steps we have $A = B_1 \ldots B_n C_n$, where C_n is the minimal tail of $B_n C_n$ with respect to B_n. The process terminates if the maximal left star of C_n is 1. The question is: Does this process always terminate?

5. Regularity of Noncounting Classes

A language $A \subset \Sigma^*$ is *noncounting* of order n, $n \geqslant 1$ iff for all $u, v \in \Sigma^*$

$$us^n v \in A \longleftrightarrow us^{n+1} v \in A.$$

It is well-known that every star-free language is noncounting, but the converse is false in general.

Let $\sim_n$ (or simply $\sim$ if n is understood) be the smallest congruence on Σ^* satisfying $s^n \sim s^{n+1}$ for all $s \in \Sigma^*$. Let $M = \Sigma^*/\sim$ and let $\mu : \Sigma^* \to M$ be the natural morphism mapping each element $s \in \Sigma^*$ into the equivalence class [s] of $\sim$ containing s. Then $m\mu^{-1}$ is the set of all words of Σ^* that are in the same equivalence class. The question is: Is $m\mu^{-1}$ a regular set for all in $m \in M$?

In case $\Sigma = \{\sigma\}$, we find:

$$[1] = 1, [\sigma] = \sigma, \ldots, [\sigma^{n-1}] = \sigma^{n-1}, [\sigma^n] = \sigma^n\sigma^*.$$

From now on assume therefore that Σ has at least two elements. The case $n = 1$, i.e. the case where M is idempotent, has been characterized by Green and Rees [15]. They have shown that $\sim_1$ is of finite index, i.e. that $M = \Sigma^*/\sim_1$ is a finite monoid. It follows that all the classes for $n = 1$ are regular, in fact star-free [3]. Thus every noncounting language of order 1 is star-free.

It is easy to show that $\sim_2$ is of infinite index in case card $\Sigma \geqslant 3$ using a result of Thue [30]. He has shown that there is an infinite set of "squareless" words, i.e. words s such that $s = ut^2v$ implies $t = 1$. This set of words is clearly

noncounting of order 2, and each word constitutes a distinct equivalence class. The same argument shows that $\sim_n$ is of infinite index for all $n \geqslant 2$ and card $\Sigma \geqslant 3$. In the case $n \geqslant 3$ one can use another result of Thue to show that $\sim_n$ is of infinite index for card $\Sigma \geqslant 2$. Thue has shown that there is an infinite number of "cubeless" words (words s such that $s = ut^3v$ implies $t = 1$) for card $\Sigma \geqslant 2$. This leaves only the case card $\Sigma = 2$ and $n = 2$, which was settled by Brzozowski, Culik and Gabrielian [3]. Let $\Sigma = \{\sigma, \tau\}$ and let $f : \Sigma^* \to \Sigma^*$ be the monoid morphism defined by

$$\sigma f = \sigma\sigma\tau \quad \text{and} \quad \tau f = \sigma\tau\tau.$$

One can show that $\sigma f^i \sim_2 \sigma f^j$ iff $i = j$. Hence $\{[\sigma], [\sigma f], \ldots, [\sigma f^i], \ldots\}$ is an infinite set of equivalence classes, and so $\sim_2$ is of infinite index. It was shown in [3] that $[\sigma f^i]$ is regular for all $i \geqslant 0$. This supports the conjecture that all classes are regular.

This problem was considered by Simon in 1970. The following is a reworking of his unpublished observations.

Let $M = \Sigma^*/\sim$ and let $r \in M$. Define the set

$$N_r = \{m \in M \mid r \notin M\, m\, M\}.$$

It is easily seen that N_r is an ideal of M, i.e. that $MN_rM = N_r$. We can construct the Rees quotient monoid [5] $M/N_r = Q_r$ of M with respect to N_r as follows. First, the Rees congruence $\leftrightarrow$ induced by N_r is defined by

$$m \leftrightarrow m' \text{ iff } m \in N_r \text{ and } m' \in N_r, \text{ or } m = m'.$$

Then $Q_r = M/N_r = M/\leftrightarrow$. Identify each congruence class of $\leftrightarrow$ containing a single element m with that element, and let 0 be a new element not in $M - N_r$ that corresponds to the class of all elements of N_r. Then we can view Q_r as consisting of $(M - N_r) \cup 0$, and having the multiplication $\circ$ defined by

$$m \circ m' = \begin{cases} mm', & \text{if } mm' \notin N_r \\ 0, & \text{otherwise.} \end{cases}$$

Note that 0 is indeed the zero element of Q_r. Let ν be the natural morphism $\nu : M \to Q_r$ mapping each element $m \in M$ into the equivalence class of $\leftrightarrow$ containing m. Let θ be the composition of μ and ν; i.e. we have

$$\Sigma^* \xrightarrow{\mu} M \xrightarrow{\nu} Q_r$$

and

$$\Sigma^* \xrightarrow{\theta = \mu\nu} Q_r.$$

We are interested in $r\mu^{-1}$. Since $r \notin N_r$, $r\nu^{-1} = r$. Thus $r\mu^{-1} = r\nu^{-1}\mu^{-1} = r(\mu\nu)^{-1} = r\theta^{-1}$, and we conclude that if Q_r is finite, then $r\mu^{-1} = r\theta^{-1}$ is regular. However, it is not known whether $\leftrightarrow$ on M is of finite index.

A second approach is as follows. Define on any monoid M the following relations:

$$m \mathrel{L} m' \text{ iff } m \in Mm' \text{ and } m' \in Mm;$$

$$m \mathcal{R} m' \text{ iff } m \in m'M \text{ and } m' \in mM;$$

$$m \mathcal{H} m' \text{ iff } m \mathcal{L} m' \text{ and } m \mathcal{R} m'.$$

These are the well-known Green equivalence relations. Let L_m, R_m and H_m denote the corresponding classes containing m. Then

$$R_m = \{m' \in M \mid m \in m'M \text{ and } m' \in mM\},$$

$$L_m = \{m' \in M \mid m \in Mm' \text{ and } m' \in Mm'\},$$

$$H_m = R_m \cap L_m.$$

If $M = \Sigma^*/\sim$ then it satisfies $m^n = m^{n+1}$ for all $m \in M$; hence M is $\mathcal{H}$-trivial; i.e. each $\mathcal{H}$-class consists of a single element. For suppose $m_1 \mathcal{H} m_2$. Then there exist $u,v \in M$ such that

$$m_1 = um_2 \text{ and } m_2 = m_1 v.$$

Thus $m_1 = um_2 = um_1 v = u^n m_1 v^n = u^n m_1 v^{n+1} = m_1 v = m_2$. Hence we have

$$m = H_m = R_m \cap L_m,$$

and

$$m\mu^{-1} = R_m\mu^{-1} \cap L_m\mu^{-1}.$$

Thus if one could prove that R_m and L_m are regular for each $m \in M$ one would have the result that $m\mu^{-1}$ is regular.

6. Optimality of Prefix Codes

Our final problem is a conjecture by Schützenberger [25].

Let Σ be a finite alphabet. A *code* C over Σ is a subset of Σ^* such that for all $s_i,t_i \in C$,

$$s_1 \ldots s_n = t_1 \ldots t_m$$

implies $n = m$ and $s_i = t_i$ for $i = 1, \ldots, n$. If C is a code and $s \in C^*$ we will say that s is a *message*. Thus every message is uniquely decipherable. For example $\{a,ab,ba\}$ is not a code for $a(ba) = (ab)a$. A code C is *prefix* iff no word of C is a prefix of any other word of C.

Two words $s,t \in \Sigma^*$ are *commutatively equivalent*, $s \sim t$, iff they differ only in the order of their letters, i.e. iff for each $\sigma \in \Sigma$

$$\binom{s}{\sigma} = \binom{t}{\sigma}.$$

Two languages $A,B \subset \Sigma^*$ are commutatively equivalent, $A \sim B$, iff there is a bijection θ from A to B such that $s \in A$ and $s\theta \in B$ are commutatively equivalent. Thus $A \sim B$ iff one can write $A = \{s_1,s_2, \ldots ,s_i, \ldots\}$ and $B = \{t_1,t_2, \ldots ,t_i, \ldots\}$ where $s_i \sim t_i$ for all i. For example let

$$C_0 = \{\sigma, \tau\sigma, \tau\tau\sigma, \tau\tau\tau\sigma, \tau\tau\tau\tau\}$$

and

$$C_1 = \{\sigma, \tau\sigma, \sigma\tau\tau, \tau\sigma\tau\tau, \tau\tau\tau\tau\}.$$

Clearly $C_0 \sim C_1$. Note that C_0 is prefix while C_1 is not.

Conjecture: Every code is commutatively equivalent to a prefix code.

The conjecture is supported by a theorem of Perrin and Schützenberger about a restricted class of codes.

Let $C \subset \Sigma^*$ be a code and $s,t \in \Sigma^*$. The pair (s,t) is *synchronizing* for C iff for any message w, if $w = ustv$ for some $u,v \in \Sigma^*$ then us and tv are also messages. In other words the appearance of st in any message w permits us to cut w into two shorter messages. For example, let

$$C_2 = \{\sigma\sigma, \sigma\tau, \tau\sigma, \tau\tau\sigma, \tau\tau\tau\}.$$

The pair $(\sigma\tau, \sigma\tau)$ is not synchronizing for C_2 because the message $\tau(\sigma\tau)(\sigma\tau)\sigma$ contains the pair but neither $\tau\sigma\tau$ nor $\sigma\tau\sigma$ are messages. However, the pair $(\sigma\tau\tau\sigma, \tau\sigma)$ can be verified to be synchronizing.

A code C has *bounded synchronization delay* if there exists an integer n such that each pair (s,t) satisfying

$$s = s_1 \ldots s_n \text{ and } t = t_1 \ldots t_n,$$

with $s_i, t_i \in C$ for $i = 1, \ldots, n$, is synchronizing. Codes with bounded delay have been studied in several papers; see [25]. One can verify that code C_1 does not have bounded delay because (τ^{4n}, τ^{4n}) is not synchronizing for any n. The code $C_4 = \{\sigma, \sigma\tau\tau, \tau\sigma, \tau\sigma\tau\tau\}$ has delay bound 1.

The main result of [25] shows that every code with bounded synchronization delay is commutatively equivalent to a prefix code. However, the general case remains open, even for regular codes.

7. Concluding Remarks

The problems described above do not appear to have any immediate direct applications to computer science. However, if one accepts the premise that mathematics should form the foundation of theoretical computer science, then the questions are quite relevant. They demonstrate our lack of proper understanding of the basic mathematical objects involved.

To end on a more positive note, I would like to mention two difficult problems that were recently solved. Since 1967 the following question was open: Given a regular language A can one decide whether there exists an integer n such that $A^n = A^*$? The problem was solved in 1978 by Hashiguchi [16] and Simon [27]; the two solutions are independent and use different techniques.

The second problem concerns the dot-depth hierarchy. Let $\mathcal{B}_0$ be the family of finite or cofinite languages. For any family $\mathcal{F}$ of languages let $\mathcal{F}M$ and $\mathcal{F}B$ denote the closure of $\mathcal{F}$ under concatenation and boolean operations respectively. Let

$$\mathcal{B}_n = \mathcal{B}_0(MB)^n.$$

Then $\mathcal{B}_n$ is the family of languages that can be expressed with n or fewer levels of concatenation, and languages in $\mathcal{B}_n - \mathcal{B}_{n-1}$ are said to be of *dot depth n*. The question whether the dot-depth hierarchy

$$\mathcal{B}_0 \subset \mathcal{B}_1 \subset \ldots \subset \mathcal{B}_n \subset$$

is infinite (i.e. whether $\mathcal{B}_{n+1} \neq \mathcal{B}_n$ for all n) was open since 1967. It was settled by Brzozowski and Knast in 1977 [4]: the hierarchy is infinite.

Acknowledgement

The author wishes to thank D. Thérien for useful comments on this paper.

References

[1] Backhouse, R.C., *Closure Algorithms and the Star Height Problem of Regular Languages*, Ph.D. Thesis, Imperial College (1974).

[2] Brzozowski, J.A. and Cohen, R., On decompositions of regular events, *J. ACM*, vol. 16, pp. 132-144 (1969).

[3] Brzozowski, J.A., Culik II, K. and Gabrielian, A., Classification of noncounting events, *J. Computer and System Sciences*, vol. 5, pp. 41-53 (1971).

[4] Brzozowski, J.A. and Knast, R., The dot-depth hierarchy of star-free languages is infinite, *J. Computer and System Sciences*, vol. 16, pp. 37-55 (1978).

[5] Clifford, A.H. and Preston, G.B., *The Algebraic Theory of Semigroups*, vol. 1, Math. Surveys 7, Amer. Math. Soc., Providence, R.I. (1961).

[6] Cohen, R.S., *Cycle Rank of Transition Graphs and the Star Height of Regular Events*, Ph.D. Dissertation, University of Ottawa, (1968).

[7] Cohen, R.S., Star height of certain families of regular events, *J. Computer and System Sciences*, vol. 4, pp. 281-297 (1970).

[8] Cohen, R.S. Techniques for establishing star height of regular sets, *Mathematical Systems Theory* vol. 5, pp. 97-114 (1971).

[9] Cohen, R.S., Rank-non-increasing transformations on transition graphs, *Information and Control*, vol. 20, pp. 93-113 (1972).

[10] Cohen, R.S. and Brzozowski, J.A., General properties of star height of regular events, *J. Computer and System Sciences*, vol. 4, pp. 260-280 (1970).

[11] Dejean, F. and Schützenberger, M.P., On a question of Eggan, *Information and Control*, vol. 9, pp. 23-25 (1966).

[12] Eggan, L.C., Transition graphs and the star height of regular events, *Michigan Math. J.*, vol. 10, pp. 385-397 (1963).

[13] Eilenberg, S., *Automata, Languages, and Machines*, vol. A, Academic Press, New York, (1974).

[14] Eilenberg, S., *Automata, Languages, and Machines*, vol. B, Academic Press, New York, (1976).

[15] Green, J.A. and Rees, D., On semigroups in which $x^r = x$, *Proc. Cambridge Philos. Soc.*, vol. 48, pp. 35-40 (1952).

[16] Hashiguchi, K., A decision procedure for the order of regular events, *Theoretical Computer Science*, vol. 8, pp. 69-72 (1979).

[17] Hashiguchi, K. and Honda, N., Homomorphisms that preserve star height, *Information and Control*, vol. 30, pp. 247-266 (1976).

[18] Henneman, W.H., *Algebraic Theory of Automata*, Ph.D. Dissertation, Massachusetts Institute of Technology, (1971).

[19] McNaughton, R., The loop complexity of pure-group events, *Information and Control*, vol. 11, pp. 167-176 (1967).

[20] McNaughton, R., The loop complexity of regular events, *Information Sci.*, vol. 1, pp. 305-328 (1969).

[21] McNaughton, R. and Papert, S., *Counter-Free Automata*, Research Monograph No. 65, The MIT Press, Cambridge, Mass., (1971).

[22] McNaughton, R. and Yamada, H., Regular expressions and state graphs for automata, *IRE Trans. Electronic Computers*, vol. EC-9, pp. 39-57 (1960).

[23] Paz, A. and Peleg, B., On concatenative decompositions of regular events, *IEEE Trans. Electronic Computers*, vol. EC-17, pp. 229-237 (1968).

[24] Pin, J.E., Sur le monoide syntactique de L^* lorsque L est un langage fini, *Theoretical Computer Science*, vol. 7, pp. 211-215 (1978).

[25] Perrin, D. and Schützenberger, M.P., Un problème élémentaire de la théorie de l'information, *Colloques Internationaux du C.N.R.S.*, no. 276 - Théorie de l'Information, pp. 249-260 (1977).

[26] Schützenberger, M.P., On finite monoids having only trivial sub-groups, *Information and Control*, vol. 8, pp. 190-194 (1965).

[27] Simon, I., Limited subsets of a free monoid, *Proceedings 19th Annual Symposium on Foundations of Computer Science*, pp. 143-150 (1978).

[28] Straubing, H., Families of recognizable sets corresponding to certain varieties of finite monoids, *J. Pure and Applied Algebra*, vol. 15, pp. 305-318 (1979).

[29] Thérien, D., *Classification of Regular Languages by Congruences*, Ph.D. Thesis, University of Waterloo, (1980).

[30] Thue, A., Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Videnskabsselskabets Skrifter*, 1, Mat.-Nat. Kl., Christiania, (1912).