REVUE FRANÇAISE D'AUTOMATIQUE INFORMATIQUE
RECHERCHE OPÉRATIONNELLE. INFORMATIQUE THÉORIQUE

JANUSZ A. BRZOZOWSKI

Hierarchies of aperiodic languages

Revue française d'automatique informatique recherche opérationnelle. Informatique théorique, tome 10, n° R2 (1976), p. 33-49

<http://www.numdam.org/item?id=ITA_1976__10_2_33_0>

R.A.I.R.O. Information Théorique
(vol. 10, n° 8, août 1976, p. 33 à 49).

HIERARCHIES OF APERIODIC LANGUAGES (*)

by JANUSZ A. BRZOZOWSKI (1)

Communiqué par J. F. PERROT

Abstract. — *In recent years, considerable attention has been given to the family of aperiodic languages, also known as star-free languages and noncounting regular languages. Several interesting subfamilies of aperiodic languages have been studied and characterized by the properties of the corresponding syntactic semigroups. The study of such families has been systematized by examining the position of each family in certain natural hierarchies. This paper gives a brief survey of results in this area.*

ACKNOWLEDGMENT

This work was done while the author at the University of Paris VI and VII, under the scientific exchange program between Canada and France. The financial assistance provided by this program is gratefully acknowledged.

Preliminary versions of this paper were given at the Séminaire P. Dubreil: Algèbre (28e année, 1974/1975, n° 19, 11 p.) and at the École de Printemps d'Informatique théorique 1975.

1. INTRODUCTION

Before considering the family of aperiodic languages (also known as star-free languages), we recall some basic notions about the more general family of regular languages (also known as rational languages). Precise mathematical definitions of the concepts mentioned in this section will be postponed. Kleene [12] related regular languages to finite automata by showing that for each regular language there is a finite automaton recognizing it, and that each language recognized by a finite automaton is regular. This correspondence between regular languages and finite automata has been extended in two directions: The first of these is practical, relating finite automata to certain types of sequential circuits; the second extension is mathematical, relating regular languages to finite semigroups. Thus the interest in finite automata and regular languages is shared by the theories of sequential circuits, formal languages, computing machine models and finite semigroups.

(*) Reçu octobre 1975.

(1) Computer Science Department University of Waterloo, Waterloo, Ontario, Canada.

The family of aperiodic languages is a proper subfamily of the family of regular languages, and is of interest from several points of view. First, it corresponds to feedback-free sequential circuits constructed with gates and set-reset flip-flops. Second, the corresponding automata are precisely the "permutation-free" automata. Third, the aperiodic languages constitue the subfamily of regular languages defined without the use of the star operator (hence "star-free"). Fourth, the corresponding semigroups are precisely the "group-free" semigroups. The basic relationship between star-free languages and group-free semigroups was shown by Schützenberger [19, 20]. Several other interesting characterizations are investigated in the monograph by McNaughton and Papert [14]; however, the above-mentioned four points of view provide ample motivation for the study of aperiodic languages.

Within the family of aperiodic languages, a number of interesting subfamilies have been studied. Among the earliest considered are the definite languages [2, 12, 17], introduced by Kleene. These are characterized by the property that the membership of a word in a definite language depends only on the length-k suffix of that word, for some integer $k \geqq 0$. This family is significant from the sequential circuit point of view, since definite languages correspond to feedback-free circuits constructed with gates and unit delays. Also there exist elegant (though somewhat more technical) characterizations of definite automata [17], and of the corresponding semigroups (discussed ater).

Left-right duality led naturally to reverse definite languages [2], where membership is determined by prefixes rather than by suffixes. The idea of testing both the prefix and suffix of a word, thus obtaining the "generalized definite" languages, is due to Ginzburg [11]. All these languages are special cases of the locally-testable languages first studied by McNaughton and Papert [14]. These are of independent interest and have appeared previously in formal language theory [7, 15]. (Thus, for example, each context-free language is a homomorphic image of a Dyck language and a certain locally testable language.)

A systematic study of subfamilies of aperiodic languages was begun by Cohen and Brzozowski [8] and continued by Brzozowski and Simon [6]. The first step consisted of the introduction of "dot-depth" as a measure of complexity of aperiodic languages. The following motivation led to these concepts. Feedback-free networks of gates, i. e. combinational circuits, constitute the simplest and degenerate forms of sequential circuits. Combinational networks are, of course, characterized by Boolean functions. This suggested that (*a*) all Boolean operations should be considered together when studying the formation of aperiodic languages from the letters of the alphabet, and (*b*) since concatenation (or "dot" operator) is linked to the sequential rather than the combinational nature of a language, the number of conca-

tenation levels required to express a given aperiodic language should provide a useful measure of complexity. (Such reasoning is not precise, and is given here only as an intuitive guide. However, subsequent studies provided considerable evidence that this is indeed a useful approach.) As it turns out, locally testable languages require only one level of concatenation, i. e. are of "dot-depth" one. A finer measure of complexity is obtained when one also takes into account the number n of factors used, considering concatenation as an n-ary operation [6]. With this refinement the families of definite, reverse definite, generalized definite and locally testable languages appear naturally in a hierarchy of families whose union is the family of languages of dot-depth one. Moreover, the families of semigroups corresponding to these families of languages appear in a rather natural hierarchy of semigroups, as we shall see later.

Simon [23, 24] showed the correspondence between certain languages of depth-one and $\mathscr{J}$-trivial monoids, thus providing a link to classical semigroup theory. This is pursued further in [4, 21].

There remains a large number of open problems, and this paper has been written the hope that it will stimulate further work in this area. The proofs of several key results stated here are quite involved and lengthy. We do not repeat them here, since it is our aim to provide a brief overview of the subject, the main results, and the open problems.

References to specific results are given in the text. For further general background on aperiodic languages see the papers [1, 5, 16, 19, 20] and the books by Eilenberg [9] and McNaughton and Papert [14].

2. NOTATION

If A is a finite, non-empty alphabet, A^+ (respectively A^*) is the free semigroup (respectively free monoid) generated by A. The empty word is denoted by 1, and Φ is the empty set. Any subset L of A^* is a language. The length of a word $w \in A^*$ is denoted by $|w|$. The cardinality of a set X is denoted card X. The symbol $\stackrel{\Delta}{=}$ means "is by definition".

Given languages $L, L' \subset A^*$, the following are also languages:

$$L \cup L' \qquad \text{(union)},$$

$$L \cap L' \qquad \text{(intersection)},$$

$$L = A^* - L \qquad \text{(complement)},$$

$$L.L' \stackrel{\Delta}{=} \{w;\ w = uu',\ u \in L,\ u' \in L'\} \qquad \text{(concatenation or product)},$$

$$L^+ \stackrel{\Delta}{=} \bigcup_{n \geqq 1} L^n \qquad \text{(the subsemigroup of } A^* \text{ generated by } L)$$

août 1976.

and

$$L^* \overset{\Delta}{=} \bigcup_{n \geqq 0} L^n = L^+ \cup 1 \qquad \text{(the submonoid of } A^* \text{ generated by } L\text{)}.$$

Let $\mathscr{U}_A$ (or simply $\mathscr{U}$ when A is understood) be the family of all languages over A. Evidently, $\mathscr{U}_A$ is a Boolean algebra under union, intersection and complement, and a monoid under concatenation.

Let $\mathscr{L}_A \overset{\Delta}{=} \{\{a\};\ a \in A\}$ and let $\mathscr{W}_A \overset{\Delta}{=} \{\{w\};\ w \in A^*\}$. Let $\mathscr{F}_A$ be the family of all finite languages, and $\mathscr{C}_A \overset{\Delta}{=} \{L \subset A^*;\ \bar{L} \in \mathscr{F}_A\}$ the family of cofinite languages.

For a given family $\mathscr{X}$ of languages, consider the following properties:

(*a*) $(L, L' \in \mathscr{X}) \Rightarrow (L \cup L' \in \mathscr{X})$;

(*b*) $(L \in \mathscr{X}) \Rightarrow (\bar{L} \in \mathscr{X})$;

(*c*) (i) $\{1\} \in \mathscr{X}$,

(ii) $L, L' \in \mathscr{X} \Rightarrow LL' \in \mathscr{X}$;

(*d*) $(L \in \mathscr{X}) \Rightarrow (L^* \in \mathscr{X})$.

It is well known [9, 12] that the family of regular or rational languages can be defined as the smallest family containing $\mathscr{L}_A$ and satisfying (*a*), (*c*) and (*d*), and that this family also satisfies (*b*).

Aperiodic languages can be defined as the smallest family containing $\mathscr{L}_A$, and satisfying (*a*), (*b*) and (*c*).

As we have said before, in the study of aperiodic languages it is useful to separate the closure under Boolean operations from the closure under concatenation. For any family $\mathscr{X} \subset \mathscr{U}$, denote by $\mathscr{X}\ B$ the Boolean algebra generated by $\mathscr{X}$, i. e. the smallest family containing $\mathscr{X}$ and satisfying (*a*) and (*b*). Similarly, $\mathscr{X}\ M$ denotes the monoid generated by $\mathscr{X}$, i. e. the smallest family containing $\mathscr{X}$ and satisfying (*c*).

3. APERIODIC LANGUAGES OVER A ONE-LETTER ALPHABET

For $A = \{a\}$, the family $\mathscr{A}_a$ of aperiodic languages is particularly simple. (We use $\mathscr{A}_a$ for $\mathscr{A}_{\{a\}}$, etc.) We have $\mathscr{L}_a M = \mathscr{W}_a = \{\{a^n\};\ n \geqq 0\}$. Define

$$\mathscr{B}_a \overset{\Delta}{=} \mathscr{L}_a M B.$$

Since $\mathscr{B}_a$ must be closed under union, it contains all finite languages, $\mathscr{B}_a \supset \mathscr{F}_a$. Closure under complementation implies $\mathscr{B}_a \supset \mathscr{C}_a$. One verifies that each cofinite language L can be written in the form

$$L = F \cup a^n a^*,$$

for some $n \geqq 0$ and some $F \in \mathscr{F}_a$. It is now clear that $\mathscr{F}_a \cup \mathscr{C}_a$ is closed under both complementation and union, i. e. that it is a Boolean algebra. Thus we conclude that

$$\mathscr{B}_a = \mathscr{F}_a \cup \mathscr{C}_a.$$

Moreover, note that concatenation of languages over a one-letter alphabet is commutative. Using this and the form (1) for cofinite languages, one verifies that $\mathscr{B}_a$ is closed under concatenation also, i. e.:

$$\mathscr{B}_a M = \mathscr{B}_a.$$

This implies that all aperiodic languages are in $\mathscr{B}_a$, i. e.:

$$\mathscr{A}_a = \mathscr{B}_a.$$

Thus *a language over a one-letter alphabet is aperiodic if, and only if, it is either finite or cofinite.*

If we start by closing $\mathscr{L}_a$ under Boolean operations first, we find

$$\tilde{\mathscr{B}}_a \overset{\Delta}{=} \mathscr{L}_a B = \{\Phi, \{a\}, a^*, a^* - a\}.$$

Next note that

$$\tilde{\mathscr{M}}_a \overset{\Delta}{=} \tilde{\mathscr{B}}_a M \supset \{\{a\}, a^*\} M = (\mathscr{L}_a \cup a^*) M = (\mathscr{W}_a \cup a^*) M$$

and

$$\mathscr{L}_a B M B = \mathscr{M}_a \mathrm{B} \supset \mathscr{F}_a \cup \tilde{\mathscr{C}}_a = \mathscr{B}_a = \mathscr{A}_a$$

because each finite and cofinite language can be expressed as an element of $(\mathscr{W}_a \cup a^*)\, MB$. Since, obviously,

$$\tilde{\mathscr{M}}_a B \subset \mathscr{A}_a,$$

we have

$$\mathscr{A}_a = \mathscr{B}_a = \tilde{\mathscr{M}}_a B.$$

These observations are summarized in Figure 1. For each inclusion, we provide an example of a language which proves the inclusion is proper.

août 1976.

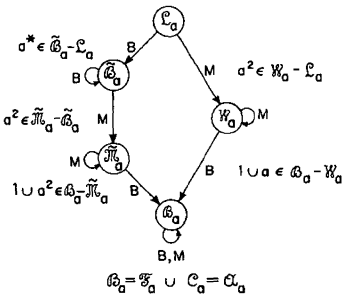

Figure 1.
Aperiodic languages over a one-letter alphabet.

4. INITIAL PHENOMENA [8]

We now assume that the alphabet A is fixed, and card $A > 1$. We use $\mathscr{L}$ for $\mathscr{L}_A$, etc. As in the one-letter case, we have $\mathscr{W} \stackrel{\Delta}{=} \mathscr{L} M$ and

$$\mathscr{B}_0 \stackrel{\Delta}{=} \mathscr{L} M B = \mathscr{F} \cup \mathscr{C}.$$

However, $\mathscr{B}_0 M \neq \mathscr{B}_0$ since (for $A = \{ a, b \}$) the language $\{ a, b \}^* a = \bar{\Phi}.a$ is in $\mathscr{B}_0^2$, but is neither finite nor cofinite. Thus we proceed to define

$$\mathscr{M}_1 \stackrel{\Delta}{=} \mathscr{B}_0 M \qquad \text{and} \qquad \mathscr{B}_1 \stackrel{\Delta}{=} \mathscr{M}_1 B.$$

We will return to these families later. For now observe that

$$\mathscr{B}_1 = (\mathscr{F} \cup \mathscr{C}) M B = (\mathscr{W} \cup \mathrm{A}^*) M B,$$

since each cofinite language can be written $L = F \cup A^n A^*$, for some $n \geqq 0$ and $F \in \mathscr{F}$. Hence L can be written as a union of products where each factor is either A^* or it is in $\mathscr{W}$.

If we close $\mathscr{L}$ under Boolean operations first, we find

$$\tilde{\mathscr{B}}_0 \stackrel{\Delta}{=} \{ L;\ L \subset A \} \cup \{ L;\ \bar{L} \subset \mathrm{A} \}.$$

Thus $\tilde{\mathscr{B}}_0$ is a finite Boolean algebra with $\mathscr{L} \cup \bar{A}$ as the set of atoms. Note that $\tilde{\mathscr{B}}_0 \supset \mathscr{L} \cup A^*$. Next

$$\tilde{\mathscr{M}}_1 \stackrel{\Delta}{=} \tilde{\mathscr{B}}_0 M \supset (\mathscr{L} \cup A^*) M = (\mathscr{W} \cup A^*) M.$$

Thus $\tilde{\mathscr{M}}_1 B \supset (\mathscr{W} \cup A^*) MB = \mathscr{B}_1$. Conversely,

$$\mathscr{B}_1 = \mathscr{L} M B M B \supset \mathscr{L} B M B = \mathscr{M}_1 B,$$

and

$$\mathscr{B}_1 = \tilde{\mathscr{M}}_1 B.$$

These properties are summarized in Figure 2. It is seen that, except for the few initial differences, it is not important whether $\mathscr{L}$ is closed under B or M first, since the two sequences coincide from $\mathscr{B}_1$ on.

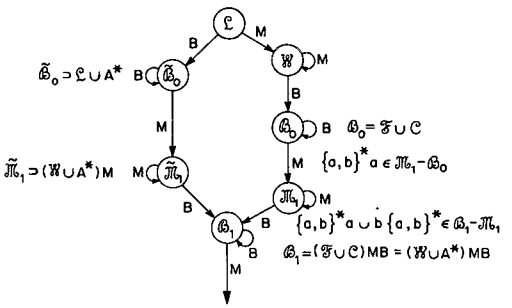

Figure 2.
Initial families for card $A \geqq 2$.

5. THE DOT-DEPTH HIERARCHY [8]

The sequence $(\mathscr{B}_i)$ of Boolean algebras, defined below, is called the *dot-depth hierarchy*. Let

$$\mathscr{B}_0 \overset{\Delta}{=} \mathscr{L} M B,$$

$$\mathscr{B}_{n+1} \overset{\Delta}{=} \mathscr{B}_n M B = \mathscr{B}_0 (M B)^n = \mathscr{L} (M B)^{n+1} \qquad \text{for} \quad n \geqq 0.$$

For each aperiodic language L, there exists $n \geqq 0$ such that $L \in \mathscr{B}_n$; hence

$$\mathscr{A} = \bigcup_{n \geqq 0} \mathscr{B}_n.$$

The "position" of a language in the dot-depth hierarchy can be used as a measure of its complexity. Define the *dot-depth* (or simply *depth*) of a language L by

$$d(\mathrm{L}) = 0 \qquad \text{if} \quad L \in \mathscr{B}_0,$$

$$d(L) = n \qquad \text{if} \quad L \in \mathscr{B}_n - \mathscr{B}_{n-1} \quad \text{for} \quad n > 0.$$

The depth $d(L)$ corresponds to the minimum number of concatenation levels that must be used to generate L from languages in $\mathscr{B}_0$. Also, $\tilde{\mathscr{B}}_0$ can be used instead of $\mathscr{B}_0$ since $\tilde{\mathscr{B}}_0 MB = \mathscr{B}_0 MB$; however, $\mathscr{B}_0$ appears to be a more natural starting point (see *Fig.* 3).

août 1976.

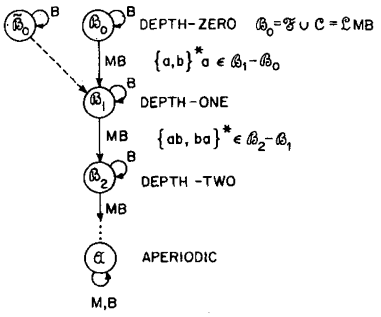

Figure 3.
The dot-depth hierarchy.

It has been shown recently by Brzozowski and Knast (*) that the dot-depth hierarchy is infinite for card $A > 1$.

For $A = \{a, b, c\}$, the language

$$L_1 = \{a, b, c\}^* b a^* = \overline{\Phi}.b.\overline{\overline{\Phi}.\{b, c\}.\overline{\Phi}}$$

is of depth 2, i. e. $L_1 \in \mathscr{B}_2 - \mathscr{B}_1$ [8]. An example over a two letter alphabet is $L_2 = \{ab, ba\}^*$.

An upper bound for $d(L)$ has been found as follows [8]. Let n be the number of states in the reduced deterministic finite automaton $\mathfrak{U}_L$ recognizing L. Let $i_a(L)$ be the number of distinct states in input column a, $a \in A$, of the state table of $\mathfrak{U}_L$. Further, let

$$i(L) = \max\{i_a(L);\ a \in A \text{ and } i_a(L) \neq n\}.$$

Then $d(L) \leqq i(L) + 1$.

This bound is met by L_1 above. On the other hand, let $L_n = a^n a^*$, $n \geqq 0$ be over $A = \{a\}$. One verifies that $i(L_n) = n - 1$, although L_n is cofinite, and $d(L_n) = 0$.

6. THE DEPTH-ONE FINITE-COFINITE HIERARCHY [6]

In the dot-depth hierarchy, $\mathscr{B}_1 = \mathscr{B}_0\, MB$, i. e. a language in $\mathscr{B}_1$ is a Boolean function of products of *any number* of factors from $\mathscr{B}_0$. Thus only one level of concatenation is required, but this concatenation is unlimited in the number of

(*) J. A. Brzozowski and R. Knast, *The Dot-Depth Hierarchy of Star-Free Languages is Infinite*, Research Report CS-76-23, Computer Science Dept., University of Waterloo, Waterloo, Ont., Canada; April, 1976.

factors. A finer measure of complexity is obtained by limiting the number of factors as follows. Let

$$\beta_n \overset{\Delta}{=} \mathscr{B}_0^n B \qquad \text{for} \quad n \geqq 1.$$

Then

$$\beta_n \subset \beta_{n+1} \qquad \text{and} \qquad \mathscr{B}_1 = \bigcup_{n \geqq 1} \beta_n.$$

A number of subfamilies of aperiodic languages that have been studied appear naturally in the sequence $\mathscr{B}_0 = \beta_1 \subset \beta_2 \subset \ldots \subset \mathscr{B}_1$ which we refer to as the *(depth-one) finite-cofinite hierarchy*. We will also need:

$$\beta_{2L} \overset{\Delta}{=} (\mathscr{F}^2 \cup \mathscr{C}\mathscr{F} \cup \mathscr{C}^2) B = (\mathscr{F} \cup \mathscr{C}\mathscr{F} \cup \mathscr{C}) B \subset \beta_2,$$
$$\beta_{2R} \overset{\Delta}{=} (\mathscr{F}^2 \cup \mathscr{F}\mathscr{C} \cup \mathscr{C}^2) B = (\mathscr{F} \cup \mathscr{F}\mathscr{C} \cup \mathscr{C}) B \subset \beta_2.$$

An alternate description of the β families is the following:

$$\beta_1 = \mathscr{B}_0 = \mathscr{W} B = (\mathscr{W} \cup A^*) B,$$
$$\beta_{2L} = (\mathscr{W} \cup A^* \mathscr{W}) B,$$
$$\beta_{2R} = (\mathscr{W} \cup \mathscr{W} A^*) B$$
$$\beta_n = (\mathscr{W} \cup A^*)^n B = (\mathscr{F} \cup \mathscr{C})^n B \qquad \text{for} \quad n \geqq 1,$$

where $A^* \mathscr{W} = \{ A^* L;\ L \in \mathscr{W} \}$, etc. These claims are easily verified. One can also show [6] that

$$(\mathscr{W} \cup A^*)^{2n+1} B \supset (\mathscr{W} \cup A^*)^{2n+2} B \qquad \text{for} \quad n \geqq 1.$$

Therefore, $\beta_{2n+2} = \beta_{2n+1}$; however, $\beta_{2n+3} \neq \beta_{2n+1}$ for all $n \geqq 1$ [23].

A language is *definite* [2, 12, 17] if, and only if, is in β_{2L}, *reverse definite* [2, 11] if, and only if, it is in β_{2R}, *generalized definite* [11, 22] if, and only if, it is in β_2, and *locally testable* [14] if, and only if, it is in β_3. The original definitions of these families of languages were somewhat different; however, the equivalence of the definitions is easily proved [6], and the present formulation appears more natural. We reconsider these families later.

The statements about the finite-cofinite hierarchy are summarized in Figure 4.

août 1976.

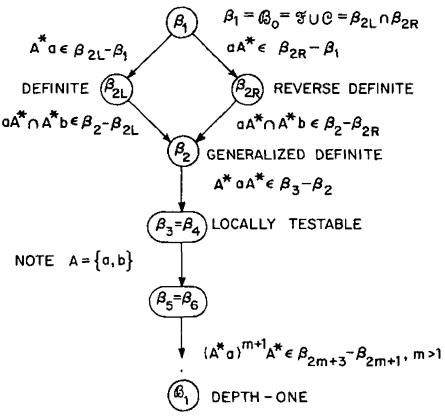

Figure 4.
The depth-one finite-cofinite hierarchy.

7. THE γ_1 HIERARCHY [23, 24]

The languages introduced here play a key role in the family of depth-one languages. We introduce a family $\alpha_{1,1}$ of languages (the reason for this notation is explained in Section 9) such that, if $L \in \alpha_{1,1}$, the membership of a word x in L can be determined solely by the set of letters appearing in x. Define

$$x\alpha \stackrel{\Delta}{=} \{ a \in A;\ x = uav \text{ and } u, v \in A^* \}$$

to be the "alphabet" of $x \in A^*$.

For $x, y \in A^*$, let $x \equiv_\alpha y$ if, and only if, $x\alpha = y\alpha$. The relation $\equiv_\alpha$ is a congruence of finite index on A^*, there being one congruence class ($[x]_\alpha$) for each subset of A. We have

$$[x]_\alpha = \Big(\bigcap_{a \in x\alpha} A^* a A^* \Big) \cap \Big(\bigcap_{a \notin x\alpha} \overline{A^* a A^*} \Big).$$

Now define $\alpha_{1,1} \stackrel{\Delta}{=} \{ [x]_\alpha;\ x \in A^* \}\ B$, and let $A^* \mathscr{L} A^* \stackrel{\Delta}{=} \{ A^* a A^*;\ a \in A \}$. One verifies that $\alpha_{1,1} = (A^* \mathscr{L} A^*)\ B$. For technical reasons, we use the family $\mathscr{G} \stackrel{\Delta}{=} A^* \cup A^* \mathscr{L} A^*$ as a generating set for $\alpha_{1,1}$. Note that $\mathscr{G}^m \subset \mathscr{G}^{m+1}$ for $m \geq 1$, and we will use the convention $\mathscr{G}^0 = \{ \Phi \}$. Let $\alpha_{m,1} \stackrel{\Delta}{=} \mathscr{G}^m B$ and $\gamma_1 \stackrel{\Delta}{=} \mathscr{G} MB$. We find $\gamma_1 = \mathscr{G} MB = \bigcup_{m \geq 0} \mathscr{G}^m B = \bigcup_{m \geq 0} \alpha_{m,1}$ and the sequence $\alpha_{0,1} \subset \alpha_{1,1} \subset \alpha_{2,1} \subset \ldots \subset \gamma_1$ will be called the γ_1 *hierarchy*.

An alternate description of $\alpha_{m,1}$ is obtained by using the "shuffle" operator $\sqcup$ [9, 10]. For $w = a_1 a_2 \ldots a_m \in A^*$,

$$A^* \sqcup w = A^* \sqcup (a_1 a_2 \ldots a_m) \stackrel{\Delta}{=} A^* a_1 A^* a_2 A^* \ldots a_m A^*$$

and

$$A^* \sqcup \mathscr{W} = \{ A^* \sqcup w;\ w \in A^* \}.$$

Let $\mathscr{W}_{\leqq m} = \{ \{ w \};\ w \in A^* \text{ and } |w| \leqq m \}$. One verifies that

$$\alpha_{m,1} = (A^* \sqcup \mathscr{W}_{\leqq m}) B, \text{ for } m \geqq 0, \text{ and } \gamma_1 = (A^* \sqcup \mathscr{W}) B.$$

Over a two-letter alphabet $A = \{ a, b \}$, the γ_1 languages can be viewed as a generalization of finite-cofinite languages [3]. Let

$$\mathscr{L}_\oplus \stackrel{\Delta}{=} \{\{a\}, \{b\}, a^+, b^+\}$$

and $\mathscr{W}_\oplus \stackrel{\Delta}{=} \mathscr{L}_\oplus M$ be the generalization of $\mathscr{L}$ and $\mathscr{W}$, respectively. Let $\mathscr{F}_\oplus$ be the closure of the family $\mathscr{W}_\oplus$ under finite unions and let $\mathscr{C}_\oplus \stackrel{\Delta}{=} \{ L;\ \bar{L} \in \mathscr{F}_\oplus \}$. Then it can be shown that

$$\gamma_1 = \mathscr{F}_\oplus \cup \mathscr{C}_\oplus = \mathscr{L}_\oplus M B.$$

Furthermore, the initial phenomena of Figure 2 have their counterpart here, for

$$\mathscr{L}_\oplus B M B = \mathscr{L}_\oplus M B M B.$$

8. THE LOCALLY-TESTABLE HIERARCHY [6, 23]

It can be shown that the membership of a word x in a locally-testable language L is determined solely by the first $k-1$ letters of x, the last $k-1$ letters of x and the set of words of length k that appear in x. Formally, $f_k(x)$ [respectively $t_k(x)$] is x, if $|x| \leqq k$, and it is the prefix (respectively suffix) of x of length k otherwise. Let

$$m_k(x) \stackrel{\Delta}{=} \{ w \in A^*;\ x = uwv \text{ and } |w| = k \}.$$

For $x, y \in A^*$ and $k > 0$, define the congruence

(*) $x \sim_k y$ if and only if,

$$f_{k-1}(x) = f_{k-1}(y),\ t_{k-1}(x) = t_{k-1}(y) \quad \text{and} \quad m_k(x) = m_k(y).$$

If $[x]_k$ is the congruence class containing x, let $\alpha_{1,k} \stackrel{\Delta}{=} \{ [x]_k;\ x \in A^* \} B$, be the family of *k-testable* languages. The reason for this notation will soon be explained. Note however that it is consistent with that of Section 7, because L is a 1-testable if, and only if, $L \in \alpha_{1,1}$. One verifies that $\alpha_{1,k} \subset \alpha_{1,k+1}$ and that $\beta_3 = \bigcup_{k \geqq 1} \alpha_{1,k}$.

août 1976.

If, in the definition (*) of $\sim_k$, we remove the condition $m_k(x) = m_k(y)$, we obtain the family of *k-generalized-definite* languages which we denote by $\alpha_{0,k}$. One verifies that $\beta_2 = \bigcup_{k \geq 1} \alpha_{0,k}$. If only t_{k-1} is tested we obtain the family of definite languages, and reverse definite languages are obtained by testing f_{k-1}. See Figure 4 for the location of these languages in the depth-one finite-cofinite hierarchy.

9. SIMON'S DEPTH-ONE HIERARCHY [23]

The membership of a word x in a language L of depth 1 can be determined by testing $f_{k-1}(x)$, $t_{k-1}(x)$ and the set $\mu_{m,k}$ of m-tuples of words of length k that "occur" in x. Thus depth-one languages are generalizations of both the k-testable and $\alpha_{m,1}$ languages; the locally testable and γ_1 hierarchies turn out to be "orthogonal".

More formally, let $W = (w_1, \ldots, w_m)$ be an m-tuple of words of length k. We say that W *occurs in* x if, and only if, there exist words $u_1, \ldots, u_m$, $v_1, \ldots, v_m$ such that $|u_1| < |u_2| < \ldots < |u_m|$ and $x = u_i w_i v_i$, for $i = 1, \ldots, m$. Let

$$\mu_{m,k}(x) = \{ W \mid W = (w_1, \ldots w_m), |w_1| = \ldots = |w_m| = k \text{ and } W \text{ occurs in } X \}.$$

By convention $\mu_{0,k} = \Phi$ for all $k \geq 1$. Note that $\mu_{m,k}(x) = \Phi$ if, and only if, $|x| < m+k-1$. For $x, y \in A^*$, $m \geq 0$, $k \geq 1$ define $x \, {}_m\!\sim_k y$ if, and only if,

(a) $$x = y \quad \text{if} \quad |x| < m+k-1$$

or

(b) $$f_{k-1}(x) = f_{k-1}(y),\ t_{k-1}(x) = t_{k-1}(y)$$

and

$$\mu_{m,k}(x) = \mu_{m,k}(y), \quad \text{otherwise.}$$

The relation ${}_m\!\sim_k$ is a congruence of finite index on A^*. Let

$$\alpha_{m,k} \triangleq \{ {}_m[x]_k;\ x \in A^* \} B.$$

One verifies that this is consistent with the previous definitions.

The hierarchy defined by ${}_m\!\sim_k$ is illustrated in Figure 5, where $\gamma_i = \bigcup_{m \geq 0} \alpha_{m,k}$ and (one verifies that) $\beta_{2m+1} = \bigcup_{k \geq 1} \alpha_{m,k}$ for $m \geq 1$ (the case of β_2 is somewhat degenerate). All the hierarchies shown are known to be infinite.

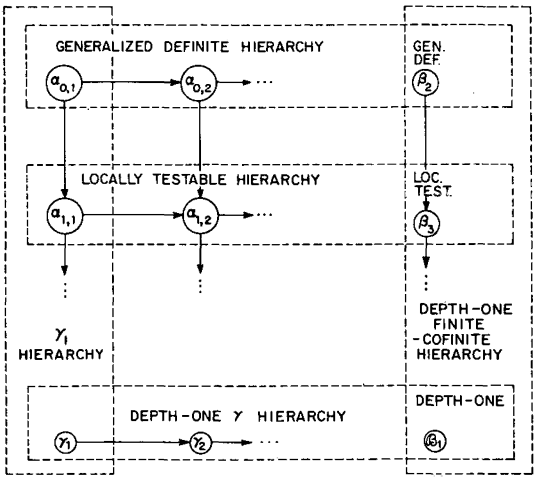

Figure 5.
Simon's depth-one hierarchy.

10. SYNTACTIC SEMIGROUPS AND MONOIDS

The congruence ${}_m\sim_k$ of the previous section can be viewed as a characterization of the family $\mathscr{B}_1$ of depth-one languages since $L \in \mathscr{B}_1$ if, and only if, there exist $m \geqq 0$ and $k \geqq 1$ such that L is a union of congruence classes of ${}_m\sim_k$. However, the problem is to decide effectively, given a regular language L, whether such m and k exist. In certain cases described below a decision procedure is available through a characterization of the syntactic semigroup or monoid of L.

For $L \subset A^*$ the *syntactic congruence*, $\equiv_L$, is defined by $x \equiv_L y$, if, and only if, for all $u, v \in A^*$, $(uxv \in L) \Leftrightarrow (uyv \in L)$. The quotient monoid $A^*/\equiv_L$ is called the *syntactic monoid* M_L of L, and $A^+/\equiv_L$ is the *syntactic semigroup*, S_L.

At first the difference between S_L and M_L appears to be rather trivial; however, it is essential in some cases to distinguish between M_L and S_L.

It is well-known that L is regular if, and only if, M_L is finite. It has been shown by Schützenberger [19, 20] that L is aperiodic if, and only if, M_L is finite and group-free (contains no groups other than the trivial one-element groups).

A number of families of languages in Simon's hierarchy have been characterized by the properties of their syntactic monoids. In this connection, the γ_1 hierarchy plays a key role. The following is known [23] :

(1) $L \in \alpha_{0,1} = \{ \varphi, A^* \}$ if, and only if, $M_L = 1$;

(2) $L \in \alpha_{1,1}$ if, and only if, M_L is finite and idempotent and commutative;

(3) $L \in \gamma_1$ if, and only if, M_L is finite and $\mathscr{J}$-trivial, i. e. for all $m, m' \in M_L$, $(M_L\, m\, M_L = M_L\, m'\, M_L)$ implies $(m = m')$.

These properties appear to carry over to the depth-one finite-cofinite hierarchy as follows:

"$L \in \alpha_{m,1}$ if, and only if, M_L has property P" seems to correspond to "$L \in \beta_{2m+1}$ if, and only if, for each idempotent $e \in S_L$ the submonoid $e\, S_L\, e$ has property P".

The following evidence supports this statement:

(1*) $L \in \beta_2$ if, and only if, S_L is finite and $e\, S_L\, e = e$ [6, 18, 25].

(2*) $L \in \beta_3$ if, and only if, S_L is finite and $e\, S_L\, e$ is idempotent and commutative [6, 13, 25, 26, 27].

(3*) If $L \in \mathscr{B}_1$, then S_L is finite and $e\, S_L\, e$ is $\mathscr{J}$-trivial [23].

As can be seen, the results are quite fragmentary, and the proofs of these results are quite complex. This approach appears to be very fruitful not only for classifying languages, but also monoids.

Simon's β hierarchy begins with generalized definite languages. Finite-cofinite, definite and reverse definite languages represent special cases, and can also be characterized by the corresponding semigroups [10, 27]. This is summarized in Figure 6.

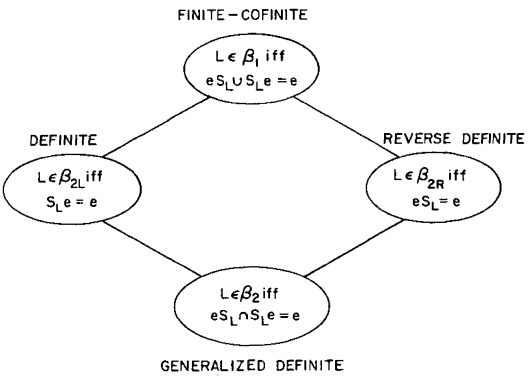

Figure 6.
Characterization of generalized definite languages.

One can generalize these ideas as follows [4]. For any monoid M and $m \in M$ define $P_m \triangleq \{ m'; m \in M\, m'\, M \}$ and $M_m \triangleq P_m^*$. Then we can consider

the family of finite monoids M in which for each idempotent e, $e M_e \cup M_e e = e$. This family is precisely the family of $\mathscr{J}$-trivial monoids. The generalization of Figure 6 is shown in Figure 7. In general, characterizations of the languages corresponding to the monoids of Figure 7 are not known, except for the $\mathscr{J}$-trivial case. However, for a two-letter alphabet these languages are generalizations of definite, reverse definite and generalized definite languages, where $\mathscr{F}_{\oplus} \cup \mathscr{C}_{\oplus}$ is used instead of $\mathscr{F} \cup \mathscr{C}$.

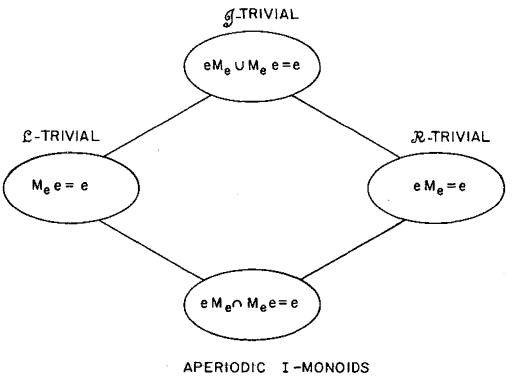

Figure 7.
Aperiodic *I*-monoids.

The correspondences between the families of languages and monoids or semigroups discussed above are examples of a more general result of Eilenberg [10]. A family of finite monoids is an *M-variety* if, and only if, it is closed under:

1) the operation of taking submonoids;

2) homomorphisms;

and

3) finite direct products.

Let $\mathscr{X}_A$ be the family of all regular languages over alphabet A (subsets of A^*), and let $\mathscr{X} = \{ \mathscr{X}_A ; A \text{ is a finite alphabet} \}$. The class $\mathscr{X}$ is called *a *-variety* if, and only if:

1) $L \in \mathscr{X}_A$ implies $A^* - L \in \mathscr{X}_A$;

2) $L, L' \in \mathscr{X}_A$ implies $L \cap L' \in \mathscr{X}_A$;

3) $L \in \mathscr{X}_A$ and $a \in A$ implies

$$\{w;\ w \in A^*,\ aw \in L\} \in \mathscr{X}_A \quad \text{and} \quad \{w;\ w \in A^*,\ wa \in L\} \in \mathscr{X}_A$$

4) If $f : B^* \to A^*$ is a homomorphism of monoids (where A and B are finite alphabets), and if $L \in \mathscr{X}_A$ then $Lf^{-1} \in \mathscr{X}_B$.

août 1976.

Thus a *-variety is closed under Boolean operations, "removal of a letter" and inverse homomorphisms.

The basic result is that to each *-variety of languages corresponds an M-variety of monoids and vice versa.

In a similar way, if we consider subsets of A^+ instead of A^*, we obtain a $^+$-variety, and, if we consider finite semigroups instead of monoids, we obtain an S-variety. Again $^+$-varieties and S-varieties correspond.

11. AUTOMATA

We have already mentioned the following correspondences:

Regular Languages $\leftrightarrow$ Finite Automata

Aperiodic Languages $\leftrightarrow$ Permutation-Free Automata

Definite Languages $\leftrightarrow$ Definite Automata

A characterization of depth-one languages in terms of automata has been found by Simon [23]. We state the result briefly.

A chain-reset is a finite automaton in which the set Q of states can be linearly ordered, say, $q_0, q_1, \ldots, q_m$, in such a way that, for each $q_i \in Q - \{ q_m \}$, the next state under any letter of the alphabet is either q_i or q_{i+1}, and for all letters the next state of q_m is q_m. Then $L \in \alpha_{m,k}$, for $m, k \geqq 1$, iff the reduced automaton $\mathfrak{U}_L$ recognizing L can be covered by a cascade product of two automata $\mathfrak{U}_1$ and $\mathfrak{U}_2$, where $\mathfrak{U}_1$ is $(k-1)$-definite and $\mathfrak{U}_2$ is a parallel connection of chain resets with at most $m+1$ states. For more details see [23].

REFERENCES

1. E. BIERMAN, *Realization of Star-Free Events*, M.A.Sc. Thesis, Department of Electrical Engineering, University of Waterloo, Waterloo, Ont., Canada, 1971.
2. J. A. BRZOZOWSKI, *Canonical Regular Expressions and Minimal State Graphs for Definite Events*, Mathematical Theory of Automata, New York, 1962, pp. 529-561, Brooklyn, Polytechnic Institute of Brooklyn, 1963 (Symposia Series, 12).
3. J. A. BRZOZOWSKI, *Run Languages, Bericht Nr.* 87, Institut für Rechner-und Programstrukturen, Gesellschaft für Mathematik und Datenverarbeitung mbH, Bonn, Germany, July 1975, 17 pp.
4. J. A. BRZOZOWSKI, *On aperiodic I-monoids*, Research Report CS-75-28, Computer Science Department, University of Waterloo, Waterloo, Ont., Canada, November 1975, 18 pp.
5. J. A. BRZOZOWSKI, K. CULIK II, and A., GABRIELIAN, *Classification of Non-counting Events*, J. Computer and System Sc., Vol. 5, 1971, pp. 41-53.
6. J. A. BRZOZOWSKI and I. SIMON, *Characterizations of Locally Testable Events*, Discrete Mathematics, Vol. 4, 1973, pp. 243-271.
7. N. CHOMSKY and M. P. SCHÜTZENBERGER, *The Algebraic Theory of Context-Free Languages*, Computer Programming and Formal Systems, edited by

P. BRAFFORT and D. HIRSCHBERG, pp. 118-161, Amsterdam, North Holland Publishing Company, 1963.

8. R. S. COHEN and J. A. BRZOZOWSKI, *Dot-Depth of Star-Free Events*, J. Computer & System Sc., Vol. 5, 1971, pp. 1-16.
9. S. EILENBERG, *Automata, Languages, and Machines*, Vol. A, New York, Academic Press, 1974 (Pure and Applied Mathematics Series, 59).
10. S. EILENBERG, *Automata, Languages and Machines*, Vol. B, New York, Academic Press, 1976.
11. A. GINZBURG, *About Some Properties of Definite, Reverse Definite and Related Automata*, I.E.E.E. Trans. Electronic Computers EC-15, 1966, pp. 806-810.
12. S. C. KLEENE, *Representation of Events in Nerve Nets and Finite Automata*, Automata Studies, edited by C.E. SHANNON and J. MCCARTHY, pp. 3-41, Princeton, Princeton University Press, 1954 (Annals of Mathematics Studies, 34).
13. R. MCNAUGHTON, *Algebraic Decision Procedures for Local Testavility*, Math. Systems Theory, Vol. 8, 1974, pp. 60-76.
14. R. MCNAUGHTON and S. PAPERT, *Counter-Free Automata*, Cambridge, The M.I.T. Press, 1971 (MIT Research Monographs, 65).
15. YU. T. MEDVEDEV, *On the Class of Events Representable in a Finite Automaton (translated from Russian)*, Sequential Machines-Selected Papers, edited by E.F. MOORE, Reading, Mass., Addison-Wesley, 1964.
16. A. R. MEYER, *A Note on Star-Free Events*, J. Assoc. Comp. Machin., Vol. 16, 1969, pp. 220-225.
17. M. PERLES, O. RABIN and E. SHAMIR, *The Theory of Definite Automata*, I.E.E.E. Trans. Electronic Computers EC-12, 1963, pp. 233-143.
18. D. PERRIN, *Sur certains semigroupes syntaxiques*, Séminaires de l'I.R.I.A. Logiques et Automates, 1971, pp. 169-177.
19. M. P. SCHÜTZENBERGER, *On Finite Monoids Having Only Trivial Sub-groups*, Inform. and Control, Vol. 8, 1965, pp. 190-194.
20. M. P. SCHÜTZENBERGER, *On a Family of Sets Related to McNaughton's L-Language*, Automata Theory, edited by E.R. CAIANIELLO, pp. 320-324, New York, Academic Press, 1966.
21. M. P. SCHÜTZENBERGER, *Sur le produit de concaténation non ambigu*, (to appear in *Semigroup Forum*).
22. M. STEINBY, *On Definite Automata and Related Systems*, Ann. Acad. Scient. Fennicae, series A.I., 1969, No. 444, 57 pp.
23. I. SIMON, *Hierarchies of Events With Dot-Depth One*, Ph.D. Thesis, Dept. of Applied Analysis & Computer Science, University of Waterloo, Waterloo, Ont., Canada, 1972.
24. I. SIMON, *Piecewise Testable Events*, 2nd GI-Professional Conference on Automata Theory and Formal Languages, Kaiserslautern, Germany, May 1975. (To appear in Lecture Notes in Computer Science, Springer-Verlag, Berlin).
25. Y. ZALCSTEIN, *Locally Testable Languages*, J. Computer and System Sc., Vol. 6, 1972, pp. 151-167.
26. Y. ZALCSTEIN, *Locally Testable Semigroups*, Semigroup Forum, Vol. 5, 1973, pp. 216-227.
27. Y. ZALCESTEIN, *Syntactic Semigroups of Some Classes of Star-Free Languages*, Automata, Languages and Programming, Proceedings of a Symposium, Rocquencourt, 1972, pp. 135-144, Amsterdam, North-Holland Publishing Company, 1973.

août 1976.